

Análisis de datos

en ciencias sociales y de la salud III

PROYECTO EDITORIAL:
Metodología de las Ciencias del Comportamiento y de la Salud

Directores:
Antonio Pardo Merino
Miguel Ángel Ruiz Díaz



Queda prohibida, salvo excepción prevista en la ley, cualquier forma de reproducción, distribución, comunicación pública y transformación de esta obra sin contar con autorización de los titulares de la propiedad intelectual. La infracción de los

derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (arts. 270 y sigs. Código Penal). El Centro Español de Derechos Reprográficos (www.cedro.org) vela por el respeto de los citados derechos.

Análisis de datos

en ciencias sociales y de la salud III

Antonio Pardo • Miguel Ángel Ruíz



Consulte nuestra página web: **www.sintesis.com**
En ella encontrará el catálogo completo y comentado

Reservados todos los derechos. Está prohibido, bajo las sanciones penales y el resarcimiento civil previstos en las leyes, reproducir, registrar o transmitir esta publicación, íntegra o parcialmente, por cualquier sistema de recuperación y por cualquier medio, sea mecánico, electrónico, magnético, electroóptico, por fotocopia o por cualquier otro, sin la autorización previa por escrito de Editorial Síntesis, S. A.

© Antonio Pardo y Miguel Ángel Ruiz

© EDITORIAL SÍNTESIS, S. A.
Vallehermoso, 34. 28015 Madrid
Teléfono 91 593 20 98
<http://www.sintesis.com>

ISBN:978-84-995894-3-5
Depósito Legal: M. 35.889-2012

Impreso en España - Printed in Spain

Índice de contenidos

Presentación	15
--------------------	----

1. Modelos lineales

Qué es un modelo lineal	20
Componentes de un modelo lineal	24
El componente aleatorio	24
El componente sistemático	24
La función de enlace	25
Clasificación de los modelos lineales	26
Cómo ajustar un modelo lineal	27
Seleccionar el modelo	27
Estimar los parámetros y obtener los pronósticos	28
Valorar la calidad o ajuste del modelo	29
Ajuste global	29
Contribución de cada variable	32
Chequear los supuestos del modelo	32
Casos atípicos e influyentes	33
Apéndice 1	
Distribuciones de la familia exponencial	35
Máxima verosimilitud	38

2. Modelos lineales clásicos

Análisis de varianza	44
Seleccionar el modelo	44
Estimar los parámetros y obtener los pronósticos	46
Valorar la calidad o ajuste del modelo	47
Chequear los supuestos	49

Análisis de covarianza	50
Lógica del análisis de covarianza	51
Seleccionar el modelo	52
Estimar los parámetros y obtener los pronósticos	53
Valorar la calidad o ajuste del modelo	54
Chequear los supuestos	54
Análisis de covarianza con SPSS	56
Cómo chequear los supuestos	56
Cómo valorar el efecto del factor	58
Pendientes de regresión heterogéneas	62
Análisis de regresión lineal	63
Seleccionar el modelo	63
Estimar los parámetros y obtener los pronósticos	64
Valorar la calidad o ajuste del modelo	66
Chequear los supuestos	68
Interacción entre variables independientes	68
Dos variables cuantitativas	69
Una variable dicotómica y una cuantitativa	71
Apéndice 2	
Elementos de un modelo lineal clásico	73

3. Modelos lineales mixtos

Efectos fijos, aleatorios y mixtos	77
Qué es un modelo lineal mixto	78
Modelos con grupos aleatorios	79
Análisis de varianza: un factor de efectos aleatorios	80
Información preliminar	83
Ajuste global	84
Significación de los efectos incluidos en el modelo	85
Estimaciones de los parámetros	86
Análisis de varianza: dos factores de efectos mixtos	88
Información preliminar	90
Ajuste global	90
Significación de los efectos incluidos en el modelo	91
Estimaciones de los parámetros	92
Comparaciones múltiples	93
Modelos con medidas repetidas	94
Estructura de los datos	95
Análisis de varianza: un factor con medidas repetidas	97
Significación de los efectos incluidos en el modelo	98
Estimaciones de los parámetros	99
Comparaciones múltiples	101

Análisis de varianza: dos factores con medidas repetidas en ambos	102
Significación de los efectos incluidos en el modelo	103
Estimaciones de los parámetros	104
Análisis de varianza: dos factores con medidas repetidas en uno	104
Significación de los efectos incluidos en el modelo	105
Estimaciones de los parámetros	106
Comparaciones múltiples	108
Análisis de los efectos simples	109
Análisis del efecto de la interacción	112
Análisis de covarianza: dos factores con medidas repetidas en uno	113
Estructura de la matriz de varianzas-covarianzas residual	116
Apéndice 3	
Elementos de un modelo lineal mixto	120
Métodos de estimación en los modelos lineales mixtos	121

4. Modelos lineales multinivel

Qué es un modelo multinivel	124
Análisis de varianza: un factor de efectos aleatorios	129
Análisis de regresión: medias como resultados	131
Análisis de covarianza: un factor de efectos aleatorios	134
Análisis de regresión: coeficientes aleatorios	136
Análisis de regresión: medias y pendientes como resultados	140
Curvas de crecimiento	146
Medidas repetidas: coeficientes aleatorios	147
Medidas repetidas: medias y pendientes como resultados	150
Apéndice 4	
El tamaño muestral en los modelos multinivel	155

5. Regresión logística (I). Respuestas dicotómicas

Regresión con respuestas dicotómicas	160
La función lineal	161
La función logística	162
La transformación logit	164
Regresión logística binaria o dicotómica	166
Una covariable (regresión simple)	167
Información preliminar	168
Ajuste global: significación estadística	170
Ajuste global: significación sustantiva	172
Pronósticos y clasificación	173
Significación de los coeficientes de regresión	176
Interpretación de los coeficientes de regresión	177

Más de una covariable (regresión múltiple)	178
Información preliminar	179
Ajuste global: significación estadística	180
Ajuste global: significación sustantiva	182
Significación de los coeficientes de regresión	182
Interpretación de los coeficientes de regresión	184
Pronósticos y clasificación	185
Covariables categóricas	187
Interacción entre covariables	190
Dos covariables dicotómicas	191
Una covariable dicotómica y una cuantitativa	194
Dos covariables cuantitativas	196
Regresión logística jerárquica o por pasos	197
Supuestos del modelo de regresión logística	203
Linealidad	203
No colinealidad	204
Independencia	205
Dispersión proporcional a la media	206
Casos atípicos e influyentes	208
Casos atípicos	208
Casos influyentes	211
Apéndice 5	
Regresión probit	212

6. Regresión logística (II). Respuestas nominales y ordinales

Regresión nominal	215
El modelo de regresión nominal	216
Una variable independiente (regresión simple)	216
Ajuste global	218
Significación e interpretación de los coeficientes de regresión	219
Más de una variable independiente (regresión múltiple)	221
Ajuste global	222
Significación e interpretación de los coeficientes de regresión	224
Pronósticos y clasificación	226
Interacción entre variables independientes	227
Regresión por pasos	228
Sobredispersión	228
Regresión ordinal	229
El modelo de regresión ordinal	229
Una variable independiente (regresión simple)	231
Ajuste global	231
Significación e interpretación de los coeficientes de regresión	232

Más de una variable independiente (regresión múltiple)	234
Interacción entre variables independientes	235
<i>Odds</i> proporcionales	235
Apéndice 6	
Funciones de enlace en los modelos de regresión ordinal	236

7. Regresión de Poisson

Regresión lineal con recuentos	240
Regresión de Poisson con recuentos	242
El modelo de regresión de Poisson	243
Una variable independiente (regresión simple)	244
Ajuste global: significación estadística	244
Ajuste global: significación sustantiva	246
Significación de los coeficientes de regresión	246
Interpretación de los coeficientes de regresión	247
Una variable independiente dicotómica	248
Una variable independiente politómica	249
Más de una variable independiente (regresión múltiple)	251
Ajuste global	251
Significación de los coeficientes de regresión	252
Interpretación de los coeficientes de regresión	253
Interacción entre variables independientes	254
Dos variables independientes dicotómicas	254
Dos variables independientes cuantitativas	256
Una variable independiente dicotómica y una cuantitativa	257
Regresión de Poisson con tasas de respuesta	258
Sobredispersión	260
Apéndice 7	
Criterios de información	261
La distribución binomial negativa y el problema de la sobredispersión	262

8. Análisis loglineal

Tablas de contingencias	266
Notación en tablas de contingencias	266
Asociación en tablas de contingencias	267
Modelos loglineales jerárquicos	269
Cómo formular modelos loglineales	269
El modelo de independencia	269
El modelo de dependencia	270
Parámetros independientes	271

Tablas multidimensionales	273
El principio de jerarquía	274
Cómo estimar las frecuencias esperadas de un modelo loglineal	276
Cómo evaluar el ajuste o la calidad de un modelo loglineal	277
Cómo seleccionar el mejor modelo loglineal	278
Cómo analizar los residuos	279
Cómo ajustar modelos loglineales jerárquicos con SPSS	281
Ajuste por pasos	282
Modelos loglineales generales	292
Cómo ajustar un modelo concreto	292
Estimaciones de los parámetros	296
Estructura de las casillas	298
Tablas incompletas	298
Ceros muestrales	299
Ceros estructurales	300
Tablas cuadradas	300
Cuasi-independencia	301
Simetría completa	304
Simetría relativa	307
Tasas de respuesta	310
Comparaciones entre niveles	314
Modelos logit	316
Una variable independiente	317
Más de una variable independiente	319
Correspondencia entre los modelos logit y los loglineales	320
El procedimiento <i>Logit</i>	324
Ajuste global: significación estadística	325
Ajuste global: significación sustantiva	327
Interpretación de los coeficientes de un modelo logit	328
Apéndice 8	
Esquemas de muestreo	331
Estadísticos mínimo-suficientes	334
Grados de libertad en un modelo loglineal	335

9. Análisis de supervivencia

Tiempos de espera, eventos, casos censurados	338
Disposición de los datos	339
Tablas de mortalidad	340
Tablas de mortalidad con SPSS	347
Cómo comparar tiempos de espera	352
El método de Kaplan-Meier	354
El estadístico <i>producto-límite</i>	355

El método de Kaplan-Meier con SPSS	357
Gráficos de los tiempos de espera	359
Cómo comparar tiempos de espera	361
Regresión de Cox	366
La ecuación de regresión	367
Impacto proporcional	368
Regresión de Cox con SPSS	369
Variables independientes categóricas	373
Regresión de Cox por pasos	375
Diagnósticos del modelo de regresión de Cox	375
Casos atípicos e influyentes	376
Residuos de Cox-Snell	376
Residuos parciales	376
Diferencia en las betas	377
Covariables dependientes del tiempo	378
Cómo crear covariables dependientes del tiempo	379
Regresión con covariables dependientes del tiempo	380
Regresión con covariables cuyos valores cambian con el tiempo	382
Apéndice 9	
Intervalos de confianza para las funciones de probabilidad, supervivencia e impacto	382
Estadístico de Wilcoxon-Gehan	384
Referencias bibliográficas	387
Índice de materias	393

Presentación

Este manual de análisis de datos es el tercer volumen de una serie dedicada a revisar los procedimientos estadísticos más utilizados en el ámbito de las ciencias sociales y de la salud.

En el primer volumen hemos incluido una revisión de las herramientas estadísticas diseñadas para describir datos y una introducción a la inferencia estadística, junto con la descripción de algunas herramientas inferenciales básicas. En el segundo volumen hemos vuelto a repasar los conceptos inferenciales básicos, particularmente en todo lo relativo al contraste de hipótesis, y hemos presentado las herramientas estadísticas diseñadas para realizar inferencias con una y dos variables; también hemos incluido en el segundo volumen los modelos de análisis de varianza más utilizados y el análisis de regresión lineal. El contenido de estos dos primeros volúmenes se ha elegido pensando en los temarios que se imparten en los diferentes grados universitarios de las disciplinas englobadas bajo la denominación de ciencias sociales y de la salud.

El propósito de este tercer volumen es ofrecer el material necesario para abordar el análisis de datos desde la perspectiva de la *modelización lineal*. Se trata de un material especialmente útil para los estudiantes de posgrado, pero también para los profesores que explican modelos lineales en esos posgrados y para los investigadores que utilizan los modelos lineales para sacar partido a sus datos.

Nuestra impresión es que el mundo de los modelos lineales es demasiado complejo para los pocos valientes investigadores aplicados que deciden acercarse a él. Incluso quienes han recibido entrenamiento para entender estos modelos y para trabajar con ellos encuentran serias dificultades, no ya solo para manejarse con soltura por las diferentes distribuciones de probabilidad y los diferentes métodos de estimación que utilizan estos modelos, sino para interpretar correctamente los resultados que se obtienen cuando los modelos se van haciendo más complejos. Nuestra intención al escribir este manual es ofrecer a los estudiantes, a los profesores y a los investigadores un material asequible y útil, es decir, un material que se pueda entender sin necesidad de tener una buena base matemática y que, al poner el énfasis en la interpretación de los resultados, pueda resultar útil a quienes, sin ser analistas expertos, se ven obligados a trabajar con este tipo de modelos.

En el Capítulo 1 explicamos qué es un modelo lineal, de qué partes consta y qué tareas es necesario llevar a cabo para poder sacar partido a una herramienta de estas características. Este capítulo también incluye una sencilla clasificación de los modelos lineales.

En el Capítulo 2 hemos incluido una revisión de algunos *modelos lineales clásicos*: el *análisis de varianza*, el *análisis de covarianza* y el *análisis de regresión lineal*. El análisis de varianza y el de regresión lineal ya los hemos tratado en el segundo volumen, pero aquí los presentamos desde la perspectiva de la modelización lineal.

Los Capítulos 3 y 4 tratan sobre los *modelos lineales mixtos*. En el Capítulo 3 abordamos los modelos de análisis de varianza y covarianza, incluidos los modelos de medidas repetidas, desde una nueva perspectiva: el enfoque *mixto*. Y en el Capítulo 4 presentamos un tipo particular de modelos mixtos que parecen haber despertado bastante interés en los últimos años: los modelos *multinivel*.

Finalmente, en los Capítulos 5 al 9 ofrecemos una revisión de los *modelos lineales generalizados*: en el Capítulo 5, el modelo de *regresión logística binaria* (para respuestas dicotómicas); en el Capítulo 6, los modelos de *regresión nominal y ordinal* (para respuestas politómicas y ordinales); en el Capítulo 7, el modelo de *regresión de Poisson* (para modelar el número de eventos); en el Capítulo 8, los modelos *loglineales* (para estudiar las pautas de asociación existentes entre un conjunto de variables categóricas); y en el Capítulo 9, el *análisis de supervivencia* (para analizar tiempos de espera en presencia de casos censurados).

Un profesional o un investigador de las ciencias sociales y de la salud no es un estadístico y, muy probablemente, tampoco pretende serlo. Consecuentemente, no necesita ser un experto en los fundamentos matemáticos de las herramientas estadísticas que aplica. Al igual que en los dos volúmenes anteriores, en la elaboración de este manual hemos pretendido ofrecer una exposición asequible de los contenidos seleccionados y hemos intentado poner el énfasis en cómo razonar para elegir el procedimiento apropiado, cómo aplicarlo con un programa informático y cómo interpretar correctamente los resultados que se obtienen. Esta es la razón que justifica que hayamos prestado más atención a los aspectos prácticos o aplicados que a los teóricos o formales, aunque sin descuidar estos últimos.

Actualmente no tiene sentido analizar datos sin el apoyo de un programa informático. Ahora bien, conviene tener muy presente que, aunque las herramientas informáticas pueden realizar cálculos con suma facilidad, todavía no están capacitadas para tomar algunas decisiones. Un programa informático no sabe si la estrategia de recogida de datos utilizada es la correcta, o si las mediciones aplicadas son apropiadas; tampoco decide qué prueba estadística conviene aplicar en cada caso, ni interpreta los resultados del análisis. Los programas informáticos todavía no permiten prescindir del analista de datos. Es el analista quien debe mantener el control de todo el proceso. El éxito de un análisis depende de él y no del programa informático. El hecho de que sea posible ejecutar las técnicas de análisis más complejas con la simple acción de pulsar un botón sólo significa que es necesario haber atado bien todos los cabos del proceso (diseño, medida, análisis, etc.) antes de pulsar el botón.

Por terminar, no podemos dejar pasar la oportunidad que nos brinda esta presentación para agradecer a nuestro compañero Ludgerio Espinosa, y a muchos de nuestros alumnos y a no pocos lectores de nuestros trabajos previos, las permanentes sugerencias hechas para mejorar nuestras explicaciones y la ayuda prestada en la caza de erratas. Los errores y deficiencias que todavía permanezcan son, sin embargo, atribuibles solamente a nosotros.

Antonio Pardo
Miguel Ángel Ruiz

1

Modelos lineales

En los procedimientos estadísticos estudiados en los dos primeros volúmenes hemos puesto todo el énfasis de nuestras explicaciones en cómo formular hipótesis referidas a problemas concretos (comparar grupos, relacionar variables) y en cómo contrastar esas hipótesis mediante alguna transformación de los datos (Z , T , X^2 , F , etc.). Hemos optado por este enfoque porque creemos que es la forma más fácil de iniciar en el análisis de datos a los estudiantes de las diferentes disciplinas del ámbito de las ciencias sociales y de la salud.

Ahora que ya estamos familiarizados con el uso de algunas herramientas estadísticas ha llegado el momento de señalar una cuestión importante: gran parte de las herramientas estadísticas que utilizamos para analizar datos son concreciones o adaptaciones de un **modelo lineal**. Aunque hasta ahora no hemos hecho referencia explícita a ello, algunos de los estadísticos más conocidos y utilizados (como, por ejemplo, la T de Student, la X^2 de Pearson o la F del análisis de varianza; estadísticos ya estudiados en los dos primeros volúmenes) se obtienen y se aplican en el marco de un modelo lineal. Los procedimientos incluidos en este volumen también representan alguna variante de un modelo lineal.

Los modelos lineales que estudiaremos comparten el objetivo de intentar describir, pronosticar o explicar el comportamiento de una variable dependiente o respuesta, o alguna transformación de la misma, mediante la combinación *lineal* de una o más variables independientes o predictoras¹. Lo que los distingue es, básicamente, la naturaleza

¹ En algunos modelos lineales (como en el análisis de correlación canónica o en el análisis multivariado de varianza) es posible incluir más de una variable dependiente, pero este tipo de modelos no serán tratados aquí. El lector interesado en ellos puede consultar Tabachnick y Fidell, 2001.

de la variable dependiente que se desea modelar (cuantitativa, dicotómica, politómica, ordinal, etc.) y, como consecuencia de ello, la distribución de probabilidad que se utiliza para representarla.

Este capítulo ofrece una descripción de las características generales de este tipo de modelos y de la forma de trabajar con ellos. En el resto de los capítulos estudiaremos varios modelos concretos. Para ampliar los conceptos que se explican aquí puede consultarse Agresti (2002, 2007), Ato y otros (2005), Brown y Prescott (1999), Dunteman y Ho (2006), Gill (2001), Harrell (2001), Hutcheson y Sofroniou (1999), McCullagh y Nelder (1989), y McCulloch y Searle (2001).

Qué es un modelo lineal

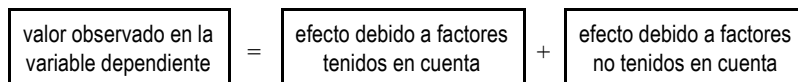
En el contexto del análisis de datos, un **modelo** es una ecuación matemática que sirve para representar de forma resumida la *relación entre dos o más variables*; el resumen de esa relación se basa en unos pocos números llamados *parámetros*.

Es posible formular muchas clases diferentes de modelos para representar la relación entre variables, pero los más simples y flexibles de todos ellos son los *lineales*. Un **modelo lineal** es una ecuación en la que los parámetros se interpretan como constantes fijas (volveremos sobre esto). En esencia, un modelo lineal intenta describir una **variable dependiente** o **respuesta** como el resultado de la combinación de un conjunto de **efectos**.

Las variables sometidas a estudio en el ámbito de las ciencias sociales y de la salud dependen, por lo general, de multitud de factores diferentes. Por tanto, cuando un sujeto obtiene una puntuación en una variable cualquiera, es realista pensar que los factores (causas) que han determinado esa puntuación son numerosos y variados; y también es realista pensar que en una investigación concreta solo será posible controlar y medir un número reducido de todos ellos.

Esta sencilla reflexión nos pone en la pista de los elementos que debe incluir un modelo que pretenda dar cuenta de la realidad; de hecho, nos permite comenzar representando la estructura de un modelo lineal tal como muestra la Figura 1.1. Un modelo lineal es, en primer lugar, un *intento de describir el valor observado en una variable dependiente o respuesta a partir del efecto debido a un conjunto de factores tenidos en cuenta y a un conjunto de factores no tenidos en cuenta*.

Figura 1.1. Estructura de un modelo lineal

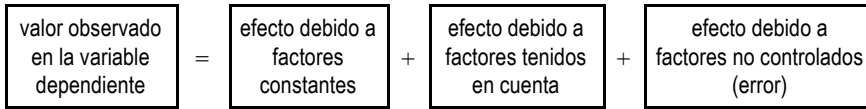


A los factores tenidos en cuenta se les suele llamar **variables independientes** o **predictoras**; son las variables explícitamente incluidas en el modelo con intención de evaluar su efecto sobre la variable dependiente.

Los factores no tenidos en cuenta son las variables cuyo efecto, aun pudiendo ser importante para describir la variable dependiente, no interesa estudiarlo o no resulta posible hacerlo. Sobre estos factores no tenidos en cuenta el investigador puede decidir ejercer o no algún tipo de **control**. Puede ejercerse control sobre una variable manteniéndola *constante* (por ejemplo, evaluando a todos los sujetos bajo las mismas condiciones ambientales se puede controlar el efecto del entorno). Sobre otros factores no se ejerce control, bien porque no se desea², bien porque no resulta posible hacerlo³. Todos los factores no controlados son los responsables de la parte de la variable dependiente que no está explicada por el conjunto de factores controlados; constituyen, por tanto, *aquello que escapa al investigador*. Para identificar al conjunto de efectos debidos a los factores no sujetos a control se suele utilizar el término **error**⁴.

Estas consideraciones permiten reformular⁵ el modelo propuesto en la Figura 1.1 tal como muestra la Figura 1.2.

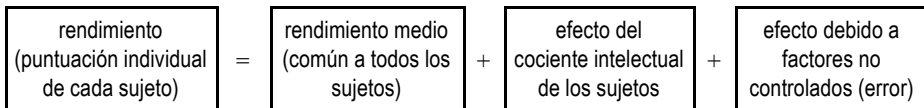
Figura 1.2. Estructura de un modelo lineal (efectos debidos a factores tenidos en cuenta desglosados)



Un ejemplo concreto puede ayudar a entender mejor la estructura de un modelo lineal. Imaginemos que estamos interesados en evaluar el efecto del cociente intelectual sobre el rendimiento académico. La Figura 1.3 muestra el resultado que se obtiene al formular este interés en el formato de un modelo lineal.

Ahora podemos dar un paso más e intentar formular matemáticamente el modelo propuesto en la Figura 1.3 (no olvidemos que un *modelo* es una *ecuación*). Esa formulación debería incluir un término para representar el rendimiento medio, uno más para

Figura 1.3. Estructura de un modelo lineal (ejemplo)



² Por ejemplo, en un estudio sobre el rendimiento académico, la *inteligencia* es un factor importante, pero el investigador puede no estar interesado en controlar su efecto, es decir, puede decidir utilizar sujetos con diferentes niveles de inteligencia simplemente porque desea que sus resultados sean más generalizables.

³ Por ejemplo, la historia individual es algo en lo que los sujetos claramente difieren pero sobre lo que un investigador no tiene, por lo general, ningún tipo de control.

⁴ El término *error* también recoge el efecto debido al hecho de que las variables que suelen utilizarse en el ámbito de las ciencias sociales y de la salud no es posible medirlas con total precisión; en los números que se analizan existe un *error de medida* implícito sobre el que no se tiene todo el control.

⁵ Judd, McClelland y Ryan (2009) resumen la estructura de un modelo lineal como *datos = modelo + error*. Con *modelo* se refieren al efecto de los factores mantenidos constantes más el efecto de los factores tenidos en cuenta.

representar el efecto del cociente intelectual y otro más para representar el error. Esto puede hacerse de diferentes formas. Una de ellas nos puede resultar bastante familiar si recordamos lo ya estudiado en el Capítulo 10 del segundo volumen a propósito del **análisis de regresión lineal**:

$$Y_i = \beta_0 + \beta_1 X_i + E_i \quad [1.1]$$

donde

Y_i = variable dependiente (rendimiento).

β_0 = efecto debido al conjunto de factores que se mantienen constantes.

$\beta_1 X_i$ = efecto debido al factor tenido en cuenta (cociente intelectual).

E_i = efecto debido al conjunto de factores no controlados (error).

(el subíndice i sirve para identificar los casos: $i = 1, 2, \dots, n$). Los términos β_0 y E_i representan el efecto debido al conjunto de factores no tenidos en cuenta. β_0 recoge el efecto debido al conjunto de factores comunes a todos los sujetos; por tanto, toma el mismo valor para todos ellos. Bajo ciertas condiciones que estudiaremos, β_0 es la media de la variable dependiente Y (la media es una forma sencilla y razonable de cuantificar la parte de la variable dependiente que comparten todos los sujetos).

El término E_i representa el efecto debido al conjunto de factores no sujetos a control: refleja la discrepancia existente entre lo que se desea explicar (Y) y lo que se consigue explicar ($\beta_0 + \beta_1 X_i$); de ahí el nombre de **error** que suele recibir. Y, dado que E_i representa justamente la parte de la variable dependiente que no explican los factores tenidos en cuenta, el modelo [1.1] suele formularse para dar cuenta, no de los valores individuales de la variable dependiente (los cuales solo pueden pronosticarse con error), sino de sus valores esperados (que representaremos mediante μ_i):

$$\mu_i = \beta_0 + \beta_1 X_i \quad [1.2]$$

Por tanto, los errores de un modelo lineal se interpretan como las desviaciones de los valores esperados de sus correspondientes observados:

$$E_i = Y_i - \mu_i \quad [1.3]$$

El término $\beta_1 X_i$ representa el efecto del factor tenido en cuenta, es decir, el efecto de la variable independiente (en el ejemplo, el cociente intelectual). β_1 es una cantidad fija que indica cómo se relaciona X (el cociente intelectual) con Y (el rendimiento académico). Cuando X es una variable cuantitativa, β_1 indica cómo cambia Y por cada unidad que cambia X . Esta propiedad del modelo es la que le confiere su principal característica: el cambio pronosticado en Y es constante (es decir, *lineal*) para cada cambio de una unidad en X .

Cuando X es una variable categórica hay que matizar el significado de β_1 . Imaginemos, por ejemplo, que el *cociente intelectual* es, en lugar de una variable cuantitativa, una variable categórica con tres niveles: 1 = “bajo”, 2 = “medio” y 3 = “alto”. Puesto

que los códigos numéricos asignados a los niveles de la variable (1, 2, 3) son arbitrarios, no tiene sentido interpretar β_1 como el cambio en Y asociado a cada unidad de cambio en X . Lo que indica β_1 es, más bien, el cambio en Y asociado al cambio de categoría o nivel en X . Y para poder reflejar esta peculiaridad se recurre a una formulación distinta de la propuesta en [1.1]:

$$Y_{ij} = \mu + \alpha_j + E_{ij} \quad [1.4]$$

(el subíndice j sirve para identificar las diferentes categorías de la variable independiente o factor: $j = 1, 2, \dots, J$). Esta formulación es la que se utiliza, por ejemplo, en los modelos de **análisis de varianza** (ver, en el siguiente capítulo, el apartado *Análisis de varianza*).

En [1.4] se está haciendo exactamente lo mismo que en [1.1]: μ equivale a β_0 y α_j equivale a $\beta_1 X_j$. Por tanto, μ (el rendimiento medio) representa el efecto debido al conjunto de factores que se mantienen constantes y α_j representa el efecto debido al factor tenido en cuenta (el cociente intelectual). Y, de acuerdo con [1.2], el valor esperado de Y se define mediante

$$\mu_{ij} = \mu + \alpha_j \quad [1.5]$$

Esta ecuación ofrece un único pronóstico por cada nivel del factor tenido en cuenta; todos los casos agrupados bajo el mismo nivel del factor reciben el mismo pronóstico; es decir, $\mu_{ij} = \mu_j$. Por tanto, $\alpha_j = \mu_{ij} - \mu = \mu_j - \mu$. Esto significa que el efecto del factor tenido en cuenta (el cociente intelectual) viene definido por las desviaciones del rendimiento medio de cada grupo respecto del rendimiento medio de todos los sujetos.

El modelo [1.2] únicamente incluye un factor tenido en cuenta (X). Incluyendo varios de estos factores ($X_1, X_2, \dots, X_j, \dots, X_p$) se obtiene la formulación convencional del **modelo lineal clásico**:

$$\mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_j X_{ij} + \dots + \beta_p X_{ip} = \beta_0 + \sum_j \beta_j X_{ij} \quad [1.6]$$

(ahora, el subíndice j se refiere a cada uno de los p factores tenidos en cuenta; por tanto, $j = 1, 2, \dots, p$). Este modelo posee una gran utilidad; a pesar de su simplicidad, es lo bastante versátil como para dar fundamento a gran parte de las técnicas de análisis de datos que se utilizan en la investigación aplicada: admite variables categóricas y cuantitativas, variables elevadas al cuadrado, términos de interacción, etc.

Pero ocurre que, para que un modelo de estas características tenga alguna utilidad, es necesario estimar los parámetros desconocidos que incluye (los coeficientes β). Y esto requiere asumir que la distribución de la variable dependiente posee ciertas características. Lo cual significa que un modelo lineal tiene dos partes: una *que se ve* y otra *que no se ve*. La parte que se ve es la propia ecuación, la cual hace explícitos los elementos que incluye el modelo y la forma en que se combinan; la parte que no se ve es la distribución de probabilidad que se asume que sigue la variable dependiente y las restricciones que se imponen sobre los elementos de la ecuación. Veamos esto con algo más de detalle.

Componentes de un modelo lineal

De las formulaciones propuestas en la Figura 1.2 y en la ecuación [1.1] se desprende que los modelos lineales de los que nos ocuparemos aquí tienen tres componentes. En este apartado ponemos nombre a esos tres componentes y aclaramos su significado.

El componente aleatorio

Este componente identifica la *variable dependiente* o *respuesta* del modelo y define una *distribución de probabilidad* para ella.

Los valores que toma la variable dependiente se consideran realizaciones concretas de una variable aleatoria que, al igual que cualquier otra variable aleatoria, tiene su propia distribución de probabilidad (que es exactamente la misma que la de los errores definidos en [1.3]). El valor de los parámetros del modelo, es decir, el valor de los coeficientes β_j , depende de cuál sea esa distribución. Y la elección de esa distribución viene condicionada, básicamente, por la naturaleza de la variable dependiente⁶.

Si la variable dependiente es cuantitativa, lo habitual es asumir que se distribuye *normalmente* con varianza constante en cada valor de X . Si la variable dependiente es dicotómica (acierto-errores, presencia-ausencia, etc.) se suele asumir que cada observación es un ensayo de Bernoulli y que el número de aciertos en n ensayos se distribuye según el modelo de probabilidad binomial. Si la variable dependiente es un recuento (número de episodios depresivos en el último año, número de accidentes de tráfico en los últimos cinco años, etc.) hay que recurrir a alguna distribución que permita trabajar con números enteros no negativos, como la distribución de Poisson.

Una misma respuesta puede modelarse de distintas maneras, pero siempre hay alguna distribución que permite modelarla mejor que las demás. Buena parte del trabajo con modelos lineales consiste en elegir la distribución de probabilidad que mejor va a conseguir modelar la respuesta que se desea analizar.

El componente sistemático

El componente sistemático contiene las *variables independientes* o *predictoras* (parte derecha de las ecuaciones [1.2], [1.5] o [1.6]). A este componente se le suele llamar **predictor lineal** (recordemos que, puesto que los coeficientes β_j se interpretan como cantidades fijas, cada variable independiente contribuye al pronóstico final con un *cambio lineal* de tamaño β_j).

El componente sistemático admite variables independientes categóricas y cuantitativas. También admite variables transformadas. Por ejemplo, podría hacerse $X_2 = X_1^2$

⁶ Las distribuciones teóricas de probabilidad también son modelos (ecuaciones). Las utilizamos, entre otras cosas, para entender mejor los datos que analizamos. Pero no todas las distribuciones son igualmente útiles: unas permiten representar los datos mejor que otras. Por ejemplo, la distribución normal refleja mejor que otras distribuciones cómo se distribuyen las puntuaciones en inteligencia. En el ajuste de modelos lineales se utilizan distribuciones de la familia exponencial: normal, binomial, Poisson, etc. (ver Apéndice 1).

para incluir en el modelo el efecto curvilíneo de X_1 ; o podría hacerse $X_3 = X_1 X_2$ para incluir en el modelo el efecto de la interacción entre las variables X_1 y X_2 .

A diferencia de la variable dependiente Y , que se considera una variable aleatoria con distribución de probabilidad conocida, las variables independientes X_j se consideran de efectos fijos. Por tanto, en un modelo de las características del propuesto en [1.6], existen tantos pronósticos μ_i distintos como valores distintos resultan de combinar los valores de las X_j . A estos valores distintos se les llama **patrones de variabilidad**.

Si todas las variables independientes son categóricas solo habrá unos pocos patrones de variabilidad y, consiguientemente, solo unos pocos pronósticos distintos; esto es lo que ocurre, por ejemplo, en los modelos de análisis de varianza. Si el modelo incluye variables independientes cuantitativas, el número de patrones de variabilidad y el de pronósticos distintos se aproximará al número de casos; esto es lo que suele ocurrir en los modelos de regresión.

La función de enlace

El tercer componente de un modelo lineal indica cómo se relacionan los componentes sistemático y aleatorio, es decir, cómo se relaciona el predictor lineal (parte derecha de la ecuación) con el valor pronosticado por el modelo (parte izquierda de la ecuación). Por tanto, la función de enlace indica qué es lo que está pronosticando exactamente el predictor lineal. La representaremos mediante $g(\mu_i)$:

$$g(\mu_i) = \beta_0 + \sum_j \beta_j X_{ij} \quad [1.7]$$

Cada una de las distribuciones elegidas para Y (normal, binomial, Poisson, etc.) contiene una función de la media que es su *parámetro natural* o *canónico* (ver Apéndice 1). En la distribución *normal* ese parámetro es la propia media; por tanto, cuando se trabaja con la distribución normal se utiliza una función de enlace **identidad**:

$$g(\mu_i) = \mu_i = \beta_0 + \sum_j \beta_j X_{ij} \quad [1.8]$$

La distribución normal y la función de enlace identidad se utilizan, por ejemplo, en el modelo de regresión lineal y en los modelos de análisis de varianza para modelar respuestas cuantitativas. Cuando se utiliza una función de enlace identidad, los pronósticos del modelo se encuentran en la misma métrica que la variable dependiente.

El parámetro natural de una distribución binomial es el **logit** de Y , es decir, el logaritmo de la *odds* de la categoría *acierto*, siendo *acierto* una cualquiera de las dos categorías de la variable dicotómica Y (la media aquí es una proporción):

$$g(\mu_i) = \log_e \left[\frac{\mu_i}{1 - \mu_i} \right] = \beta_0 + \sum_j \beta_j X_{ij} \quad [1.9]$$

La función de enlace *logit* es útil para modelar una variable dependiente que toma valores comprendidos entre 0 y 1 como, por ejemplo, una probabilidad. La distribución

binomial y la función de enlace logit se utilizan en el análisis de regresión logística (binaria, nominal, ordinal) y en el análisis de los tiempos de supervivencia. Puesto que un modelo de regresión logística pronostica el logit de Y , los pronósticos y la variable dependiente están en métricas distintas.

El parámetro natural de una distribución de Poisson es el **logaritmo** de la media (la media aquí es un número entero no negativo):

$$g(\mu_i) = \log_e(\mu_i) = \beta_0 + \sum_j \beta_j X_{ij} \quad [1.10]$$

La distribución de Poisson y la función de enlace *logarítmica* son útiles para modelar frecuencias; permiten obtener una buena representación de las variables cuyos valores son enteros no negativos. Se utilizan en la regresión de Poisson y en los modelos loglineales. Los pronósticos y la variable dependiente están en métricas distintas.

Clasificación de los modelos lineales

Atendiendo a las características de sus componentes podemos comenzar distinguiendo tres tipos de modelos lineales:

1. **Modelos lineales clásicos:** análisis de regresión, análisis de varianza, análisis de covarianza (todos ellos englobados en lo que se conoce como *modelo lineal general*; no confundir *general* con *generalizado*). Son útiles para modelar respuestas cuantitativas. Utilizan una función de enlace *identidad*.
2. **Modelos lineales mixtos:** análisis de varianza de efectos aleatorios y de efectos mixtos, análisis de regresión multinivel. Al igual que los modelos clásicos, también sirven para modelar respuestas cuantitativas y utilizan una función de enlace identidad, pero se diferencian de ellos en que pueden incluir más de un término error (no requieren asumir varianzas iguales ni observaciones independientes).
3. **Modelos lineales generalizados:** regresión logística, regresión de Poisson, regresión nominal, regresión ordinal, modelos loglineales y logit, modelos de impacto proporcional, etc. Sirven para modelar respuestas no cuantitativas (respuestas dicotómicas, politómicas, ordinales, frecuencias) y respuestas cuantitativas que no pueden analizarse con los modelos clásicos o mixtos (como los tiempos de supervivencia). Utilizan, básicamente, las funciones de enlace logit y logarítmica.

Todos ellos son, tal como argumentan Nelder y Wedderburn en su influyente artículo de 1972, modelos lineales *generalizados* (ver también Gill, 2001), pero clasificarlos como *clásicos*, *mixtos* y *generalizados* nos ayudará a identificarlos más fácilmente.

Aunque la mayoría de los modelos que estudiaremos están diseñados para analizar *una sola variable dependiente*, también hay modelos que permiten analizar simultáneamente *más de una variable dependiente*, como el análisis de correlación canónica y el análisis multivariado de varianza (estos modelos no serán tratados aquí; ver, por ejemplo, Tabachnick y Fidell, 2001). Y también hay modelos lineales que, aunque en su formulación contienen variables a ambos lados de la ecuación, se utilizan para explorar

posibles pautas de asociación sin distinguir entre variables independientes y dependientes. Tal es el caso de los modelos loglineales que estudiaremos en el Capítulo 8 y del análisis factorial, exploratorio y confirmatorio, que no estudiaremos aquí porque ya se ha tratado en otro volumen de la colección (ver Abad, Olea, Ponsoda y García, 2011).

Cómo ajustar un modelo lineal

Aunque cada uno de los modelos que estudiaremos posee sus propias peculiaridades, ajustar modelos lineales requiere, por lo general, realizar cuatro tareas:

1. Seleccionar el modelo que podría dar cuenta de la relación estudiada.
2. Estimar los parámetros del modelo y obtener los pronósticos.
3. Evaluar la calidad o ajuste del modelo.
4. Chequear los supuestos y la posible presencia de casos atípicos e influyentes.

En la práctica, el ajuste de modelos lineales no consiste en aplicar estas tareas de forma secuencial y en un solo paso. Según veremos, el ajuste de modelos lineales suele ser un proceso cíclico que requiere volver atrás una y otra vez incorporando las modificaciones que va sugiriendo el análisis hasta llegar al modelo final.

Seleccionar el modelo

Cuando se decide utilizar un modelo lineal, la primera tarea que hay que abordar es la de elegir el *tipo* de modelo (clásico, mixto, generalizado) más apropiado para analizar los datos disponibles. En esta elección, el criterio determinante suele ser el tipo de variable dependiente que se desea modelar (ver apartado anterior).

Cualquiera que sea el tipo de modelo elegido, siempre existe un modelo *nulo* y un modelo *saturado* que representan los dos extremos de un conjunto de posibilidades.

El **modelo nulo** incluye un único parámetro: el término constante β_0 . Por tanto, ofrece el mismo pronóstico para todos los casos. Toda la variabilidad de Y está representada por el término error. Puesto que no incluye ninguna variable independiente, lo consideramos el *peor* modelo posible en el sentido de que, de todos los modelos que podrían formularse, es el que menos ayuda a entender o explicar el comportamiento de la variable dependiente. No obstante, justamente por tratarse del peor modelo posible, sirve de referente con el que comparar otros modelos.

El **modelo saturado** incluye tantos parámetros como observaciones. Por tanto, el componente sistemático permite dar cuenta de toda la variabilidad de Y . Es un modelo que ofrece pronósticos perfectos, pero carece de utilidad porque no resume la información contenida en los datos. No obstante, puesto que ofrece pronósticos perfectos, sirve, al igual que el modelo nulo, como referente con el que comparar otros modelos.

El resto de modelos se encuentran entre el *nulo* y el *saturado*; todos ellos incluyen más términos (parámetros) que el nulo y menos que el saturado. Uno de esos modelos

será el que interesará formular y ajustar en cada situación concreta para valorar si consigue o no dar cuenta de la relación estudiada.

Para encontrar ese modelo pueden seguirse dos estrategias alternativas: (1) si se tiene una hipótesis concreta, es decir, una idea previa acerca de la pauta de relación estudiada, lo razonable será formular y ajustar el modelo lineal que permita contrastar esa hipótesis; (2) si no se tiene una hipótesis concreta, será preferible proceder por pasos, añadiendo o quitando términos, hasta encontrar el modelo capaz de describir la relación subyacente de la mejor forma posible. Veremos cómo aplicar ambas estrategias con cada uno de los modelos lineales que estudiemos.

Aunque la elección de un modelo lineal concreto es una tarea tanto más compleja cuanto mayor es el número de variables independientes involucradas, el objetivo de la elección siempre es el mismo: encontrar el modelo que, además de tener algún significado teórico, guarde un buen equilibrio entre dos criterios que apuntan en direcciones opuestas: (1) ser lo *bastante complejo* como para posibilitar un buen ajuste a los datos (criterio de **máximo ajuste**) y, al mismo tiempo, (2) lo *bastante simple* como para ser fácilmente interpretable y lo más generalizable posible (criterio de **parsimonia**).

Estimar los parámetros y obtener los pronósticos

Los parámetros de un modelo lineal son, por lo general, valores desconocidos que es necesario estimar a partir de los datos. Para obtener estas estimaciones pueden utilizarse diferentes criterios, pero lo habitual es elegir para β_j los valores $\hat{\beta}_j$ que consiguen hacer que los *pronósticos* del modelo se parezcan lo máximo posible a los valores *observados*.

Para esto suelen utilizarse dos métodos distintos: (1) el de **mínimos cuadrados**, que consiste en elegir para $\hat{\beta}_j$ los valores que minimizan la suma de las diferencias al cuadrado entre los valores observados y los pronosticados (ver, por ejemplo, Montgomery, Peck y Vining, 2001, págs. 14-22 y 71-82) y (2) el de **máxima verosimilitud**, que consiste en elegir para $\hat{\beta}_j$ los valores que maximizan la probabilidad de obtener los datos de hecho obtenidos (ver, por ejemplo, Dunteman y Ho, 2006, págs. 23-31).

Con los modelos clásicos ambos métodos ofrecen las mismas estimaciones. Y si se cumplen los supuestos del modelo (ver más adelante), tanto los estimadores mínimo-cuadráticos como los máximo-verosímiles son insesgados y eficientes. No obstante, con los modelos lineales clásicos suele utilizarse el método de mínimos cuadrados mientras que con los modelos mixtos y generalizados se utiliza el método de máxima verosimilitud (ver Apéndice 1). Obtener las estimaciones máximo-verosímiles de un modelo generalizado no es una tarea rápida ni sencilla⁷. De hecho, el sistema de ecuaciones que

⁷ El sistema de ecuaciones que se utiliza para obtener las estimaciones de los parámetros de la mayoría de los modelos generalizados no puede resolverse analíticamente (ver McCulloch y Searle, 2001, pág. 142). Las estimaciones de máxima verosimilitud se obtienen aplicando algoritmos de cálculo iterativo. Nelder y Wedderburn (1972) han propuesto un algoritmo llamado *mínimos cuadrados ponderados iterativamente* (basado en el método de tanteo - *scoring* - de Fisher y en el algoritmo de Newton-Raphson) que permite obtener las estimaciones de máxima verosimilitud de cualquier modelo lineal en el que se asuma para el componente aleatorio una distribución de la familia exponencial (ver, por ejemplo, Gill, 2001, págs. 39-51).

hay que resolver requiere utilizar métodos especiales de cálculo iterativo. No obstante, los programas informáticos de uso más extendido tienen resuelto este problema; todos ellos incorporan algoritmos que permiten estimar los parámetros de cualquiera de los modelos lineales que estudiaremos.

Una vez estimados los parámetros del modelo, ya es posible obtener los pronósticos $g(\hat{\mu}_i)$ que se derivan del mismo:

$$g(\hat{\mu}_i) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_p X_{ip} \quad [1.11]$$

Este modelo es idéntico al de regresión lineal ya estudiado en el Capítulo 10 del segundo volumen con otra notación ($\hat{Y}_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + \cdots + B_p X_{ip}$) y con función de enlace *identidad*.

Si los pronósticos no están en la misma métrica que la variable dependiente (es decir, si $g(\hat{\mu}_i) \neq \hat{\mu}_i$, cosa que ocurre siempre que se ajusta un modelo generalizado), hay que devolverlos a su métrica original. Y esto, no solo para obtener los pronósticos, sino para poder proceder a valorar la calidad del modelo propuesto y para realizar algunos diagnósticos.

Valorar la calidad o ajuste del modelo

Tanto el criterio de mínimos cuadrados como el de máxima verosimilitud permiten encontrar el modelo que mejor describe o resume los datos. Pero el hecho de que un determinado modelo sea el *mejor* no implica que sea *bueno*. En realidad, el mejor modelo posible puede ir desde muy malo a excelente. Esto puede apreciarse fácilmente al ajustar una recta a una nube de puntos: aunque la recta mínimo-cuadrática sea el mejor resumen de una nube de puntos, la calidad de ese resumen dependerá del grado de dispersión de los puntos en torno a la recta.

Esta reflexión debe servirnos para reparar en la importancia de detenerse a valorar la calidad del modelo elegido. Y la calidad de un modelo viene dada, básicamente, por el grado de ajuste del modelo, es decir, por el grado de parecido existente entre los valores observados y los pronosticados. Al valorar el ajuste de un modelo lineal hay que considerar dos aspectos: (1) el ajuste global del modelo y (2) la contribución individual de cada variable independiente al ajuste global.

Ajuste global

Valorar el ajuste de un modelo lineal requiere prestar atención a dos tipos de significación. Por un lado, el estudio de la **significación estadística** sirve para dar respuesta a preguntas del tipo: ¿ofrece el modelo propuesto mejor ajuste (mejores pronósticos) que el modelo que no incluye ninguna de las variables independientes elegidas? Por otro, el estudio de la **significación sustantiva** sirve para dar respuesta a preguntas del tipo: ¿consigue el modelo propuesto explicar una parte relevante o importante de la variable dependiente?

Para responder a estas preguntas es común utilizar estadísticos de **ajuste global**. En concreto, para valorar la *significación estadística* suele utilizarse un estadístico llamado **desvianza** (*deviance*; se representa mediante $-2LL$). Para valorar la *significación sustantiva* es habitual utilizar estadísticos que intentan cuantificar la **proporción de varianza común o explicada** (estadísticos como el coeficiente de determinación; suelen representarse mediante R^2).

La *desvianza* adopta diferentes formatos dependiendo del tipo de modelo lineal elegido, pero siempre representa una *cuantificación de la discrepancia existente entre los valores observados y los pronosticados*. Por tanto, la desvianza alcanza su valor máximo con el modelo *nulo* y su valor mínimo con el modelo *saturado*.

Cuando las estimaciones se basan en el método de mínimos cuadrados, la desvianza se obtiene sumando las diferencias al cuadrado entre los valores observados y los pronosticados (este estadístico ya lo conocemos con el nombre de *suma de cuadrados error* o *residual*; ver Capítulo 10 del segundo volumen). Cuando las estimaciones se basan en el método de máxima verosimilitud, la desvianza se obtiene (ver, por ejemplo, Duntelman y Ho, 2006, págs. 31-32; o Gill, 2001, págs. 56-58) comparando *dos funciones de verosimilitud en escala logarítmica*: la del modelo propuesto (LL_M) y la del modelo saturado (LL_S):

$$-2LL_M = -2(LL_M - LL_S) \quad [1.12]$$

(aunque utilizaremos con frecuencia este estadístico, no será necesario calcularlo a mano; los programas informáticos tienen resuelto esto). Puesto que la verosimilitud del modelo saturado se corresponde con el máximo ajuste posible (el modelo saturado siempre ofrece pronósticos perfectos), el resultado de la ecuación [1.12], es decir, la desvianza, está reflejando el grado en que el modelo propuesto se aleja del ajuste perfecto. En algunos modelos lineales, $-2LL_M$ se aproxima a la distribución ji-cuadrado con $n - k$ grados de libertad (n es el número de observaciones; k es el número de parámetros en que difieren el modelo saturado y el modelo propuesto). Por tanto, la desvianza ($-2LL_M$) y la distribución ji-cuadrado pueden utilizarse para valorar el ajuste global de un modelo concreto mediante el contraste de la hipótesis nula de que los parámetros extra que contiene el modelo saturado valen cero.

El rechazo de esta hipótesis estaría indicando que el modelo saturado contiene términos que mejoran significativamente el ajuste del modelo propuesto. Pero el hecho de que un determinado modelo no consiga un ajuste perfecto no significa que no pueda estar contribuyendo a mejorar nuestro conocimiento de la variable dependiente. Esto debe valorarse comparando el ajuste que consigue ese modelo con el ajuste que consigue el modelo nulo, es decir, valorando la significación estadística de los términos extra que incluye el modelo propuesto respecto del modelo nulo, lo cual equivale a contrastar la hipótesis nula de que los coeficientes extra que incluye el modelo propuesto valen cero:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0 \quad [1.13]$$

Para contrastar esta hipótesis se suele utilizar un estadístico llamado **razón de verosimilitudes** (G^2). Este estadístico se basa en las desvianzas de los dos modelos involu-

crados: el modelo nulo o modelo 0, que afirma que la hipótesis nula propuesta en [1.13] es cierta; y el modelo propuesto o modelo 1, que afirma que la hipótesis propuesta en [1.13] es falsa:

$$H_1: \beta_j \neq 0, \text{ para algún } j \quad [1.14]$$

Puesto que la desviación del modelo nulo ($-2LL_0$) refleja el máximo grado posible de desajuste (el desajuste que se obtiene al pronosticar la variable dependiente sin otra información que la propia variable dependiente), la diferencia entre esa desviación y la del modelo propuesto ($-2LL_1$) estará reflejando en qué medida el modelo propuesto consigue reducir el desajuste del modelo que *peor* ajusta.

Cuando un modelo incluye todos los términos de otro modelo más alguno adicional (a dos modelos que cumplen esta condición se les llama *jerárquicos* o *anidados*), es posible valorar la significación estadística de los términos extra que incluye el primer modelo comparando las desviaciones de ambos modelos. Por tanto, los términos extra que incluye el modelo que se desea ajustar (modelo 1) respecto del modelo nulo (modelo 0), que son justamente los términos que se están igualando a cero en la hipótesis [1.13], pueden evaluarse mediante⁸:

$$G^2_{0-1} = -2LL_0 - (-2LL_1) \quad [1.15]$$

La distribución muestral de G^2 se aproxima a la distribución ji-cuadrado con los grados de libertad resultantes de restar el número de parámetros de ambos modelos. La aproximación es tanto mejor cuanto mayor es el número de observaciones.

El rechazo de la hipótesis [1.13] indica que el modelo propuesto contribuye a reducir significativamente el desajuste del modelo nulo; o, de otro modo, que el modelo propuesto contribuye a mejorar significativamente el ajuste del modelo nulo. Ahora bien, decidir que una *mejora es estadísticamente significativa* no implica que se trate de una *mejora relevante*. Para poder afirmar esto último hay que valorar, no la significación estadística, sino la **significación sustantiva**. Y esto requiere utilizar otro tipo de estadísticos.

El estadístico habitualmente utilizado para valorar la significación sustantiva de un modelo lineal es el coeficiente de determinación (el cuadrado del coeficiente de correlación de Pearson, R^2), el cual indica, entre otras cosas, en qué proporción se consigue reducir el desajuste del modelo nulo, es decir, en qué medida se consiguen reducir los errores de predicción (las diferencias entre los valores observados y los pronosticados):

$$R^2 = \frac{-2LL_0 - (-2LL_1)}{-2LL_0} \quad [1.16]$$

⁸ Para contrastar la hipótesis [1.13] en el contexto de un modelo lineal clásico se utilizan **estadísticos F** que comparan diferentes fuentes de variabilidad, lo cual no es otra cosa que comparar desviaciones. En regresión lineal, por ejemplo, la suma de cuadrados debida a la regresión es $-2LL_0 - (-2LL_1)$ y la suma de cuadrados error es $-2LL_1$. El estadístico F es el cociente entre ambas sumas de cuadrados (es decir, entre ambas desviaciones), después de dividir cada una de ellas entre sus correspondientes grados de libertad.

Puesto que la desvianza de un modelo indica el grado de desajuste del mismo, la diferencia entre la desvianza del modelo nulo y la del modelo propuesto (es decir, G^2_{0-1}) representa la diferencia en el desajuste de ambos modelos. Dividiendo esta diferencia entre la desvianza del modelo nulo se obtiene la *proporción en que el modelo propuesto consigue reducir el desajuste del modelo nulo* (es decir, la proporción en que el modelo propuesto consigue reducir los errores de predicción del modelo nulo).

Según veremos, cuando la variable dependiente es categórica también es posible valorar la significación sustantiva de un modelo mediante el porcentaje de casos correctamente clasificados, es decir, mediante el porcentaje de pronósticos correctos (esto es algo que no tiene sentido con respuestas cuantitativas, donde un pronóstico muy parecido al valor observado, pero no idéntico, no representa un error equivalente a pronosticar, por ejemplo, “recuperado” a un sujeto “no recuperado”).

Contribución de cada variable

El hecho de que un modelo concreto esté contribuyendo a reducir el desajuste del modelo nulo no implica que todas las variables independientes o predictoras incluidas en el modelo estén contribuyendo a reducir el desajuste en la misma medida. De hecho, no es infrecuente encontrar que algunas de las variables incluidas en un modelo no contribuyen en absoluto a reducir el desajuste. Y el criterio de *parsimonia* exige eliminar del modelo todo lo irrelevante, es decir, todo aquello que no contribuya a mejorar su calidad.

Acabamos de ver que la razón de verosimilitudes definida en [1.15] sirve para valorar la significación estadística de los términos en que difieren dos modelos cuando los términos que incluye uno de ellos es un subconjunto de los que incluye el otro. Pues bien, cuando los modelos que se comparan difieren en un único término, la razón de verosimilitudes permite valorar la significación estadística de ese término. Y la significación sustantiva de un término concreto puede valorarse a partir del incremento en R^2 que produce su incorporación al modelo.

Las variables cuyos coeficientes no son significativamente distintos de cero pueden eliminarse del modelo (haciendo el modelo más simple) sin pérdida de ajuste, es decir, sin que ello afecte al valor de R^2 .

Chequear los supuestos del modelo

Ya hemos señalado que un modelo lineal tiene dos partes: la que se ve y la que no se ve. La parte que se ve es la ecuación; la parte que no se ve son los *supuestos*, es decir, las condiciones que deben darse (y que es necesario hacer explícitas) para que la ecuación propuesta funcione bien.

La calidad de un modelo estadístico tiene mucho que ver con el cumplimiento de los supuestos en los que se basa. El incumplimiento de éstos puede llevar a estimaciones sesgadas y poco eficientes, y éstas a inferencias incorrectas.

Al estudiar el modelo de regresión lineal (ver Capítulo 10 del segundo volumen) hemos hecho referencia a cinco supuestos: linealidad, no-colinealidad, independencia, homocedasticidad y normalidad. Lo dicho allí sobre el significado de cada supuesto sigue siendo válido aquí (en caso necesario, revisar el apartado *Supuestos del modelo de regresión lineal* del mencionado capítulo).

Al ajustar modelos lineales clásicos, los cuatro primeros supuestos son necesarios para que los coeficientes del modelo sean estimadores insesgados y eficientes de sus respectivos parámetros; y el supuesto de normalidad permite contrastar hipótesis sobre los coeficientes de regresión y construir intervalos de confianza.

En los modelos lineales mixtos se pueden relajar dos de los cinco supuestos de los modelos clásicos. En primer lugar, no es necesario trabajar con observaciones independientes; los modelos mixtos permiten definir diferentes estructuras de covarianza para poder modelar datos que no son independientes entre sí. En segundo lugar, no es necesario asumir que la varianza de la variable dependiente (o la varianza de los errores, que es la misma) es constante para cada patrón de variabilidad; los modelos mixtos permiten trabajar con varianzas heterogéneas.

Los modelos generalizados también permiten relajar dos de los cinco supuestos de los modelos clásicos. En primer lugar, no es necesario asumir que el componente aleatorio se distribuye normalmente: en los modelos generalizados se utilizan distribuciones de la familia exponencial distintas de la normal (binomial, binomial negativa, Poisson, etc.). En segundo lugar, no es necesario asumir homocedasticidad: aunque la media y la varianza de una distribución normal son independientes (de ahí que el supuesto de homocedasticidad típico de los modelos clásicos lleve asociado el de normalidad), esto no ocurre en el resto de distribuciones de la familia exponencial; de hecho, en las distribuciones exponenciales no-normales, el tamaño de la varianza depende del tamaño de la media (ver, en el Apéndice 1, el apartado *Distribuciones de la familia exponencial*).

El hecho de que la varianza de una distribución exponencial no-normal dependa del tamaño de su media obliga a chequear la posible presencia de *sobredispersión* cuando se trabaja con este tipo de distribuciones. La sobredispersión se da cuando la varianza observada es mayor que la esperada de acuerdo con la distribución teórica utilizada. También puede ocurrir que la varianza sea menor que la media (*infradispersión*), pero esto es más bien infrecuente.

Casos atípicos e influyentes

Valorar la calidad de un modelo estadístico requiere, finalmente, vigilar algunos detalles que podrían estar distorsionando los resultados del análisis. Estos detalles se refieren básicamente a la posible presencia de casos *atípicos* e *influyentes*.

Un **caso atípico** es un caso inusual, un caso que no se parece a los demás. Un caso puede ser atípico en la variable dependiente Y , en la(s) independiente(s) X_j , o en ambas. Los casos atípicos en Y pueden detectarse analizando los *residuos*, es decir, las diferencias entre los valores observados y los pronosticados por el modelo (los residuos son la versión muestral de los errores poblacionales definidos en [1.3]). Un residuo

excesivamente grande delata un caso mal pronosticado; es decir, un caso cuyo valor en Y se aleja de lo que cabría esperar de él de acuerdo con sus valores en las X_j . Y un caso mal pronosticado suele ser un caso atípico en Y . Según veremos, existen diferentes formas de calcular los residuos y diferentes formas de transformarlos para facilitar su interpretación (los residuos tienen otras utilidades que tendremos ocasión de ir descubriendo).

Para detectar casos atípicos en las X_j suele utilizarse un estadístico llamado *influencia* (*leverage*). Este estadístico refleja el grado de alejamiento de cada caso respecto del centro de su distribución, es decir, el grado de alejamiento del conjunto de puntuaciones de un caso respecto de las puntuaciones medias de todos los casos.

Por último, conviene tener presente que, aunque todos los casos contribuyen a estimar los parámetros de un modelo, no todos lo hacen en la misma medida. Los casos **influyentes** son casos que afectan de forma importante a los resultados del análisis. Un caso influyente no debe confundirse con un caso atípico. Los casos atípicos son casos que conviene revisar, pero no necesariamente son casos influyentes. Para que un caso pueda ser etiquetado de influyente, además de ser atípico, debe alterar de forma importante los resultados del análisis. Para detectar casos influyentes se suelen utilizar estadísticos que permiten comparar lo que ocurre cuando se incluyen todos los casos en el análisis con lo que ocurre al eliminar cada caso. Para obtener estos estadísticos se estiman $n + 1$ ecuaciones: una basada en todos los casos y las n restantes eliminando un caso cada vez. Y el diagnóstico se centra en valorar cómo van cambiando los resultados del análisis (los coeficientes del modelo, los pronósticos, los residuos) al ir eliminando cada caso.

Resumiendo

Analizar datos mediante el ajuste de un modelo lineal es un proceso que se desarrolla en fases y de forma cíclica.

El objetivo del análisis es encontrar el modelo capaz de representar o resumir de la mejor manera posible la relación existente entre las variables estudiadas. Para ello, se comienza eligiendo el tipo de modelo más apropiado para representar la variable dependiente que se tiene intención de modelar (clásico, mixto, generalizado) y las variables independientes que incluirá; la elección del tipo de modelo depende básicamente de las características de la variable dependiente; la elección de las variables independientes suele hacerse a partir de la evidencia previa disponible. A continuación se estiman los parámetros del modelo y se obtienen los pronósticos que se derivan del mismo. Tras esto, se realiza una valoración del modelo para decidir si contribuye o no a mejorar nuestra comprensión del fenómeno estudiado. En caso necesario, se eliminan los elementos inservibles del modelo y se vuelven a estimar los parámetros y a valorar la calidad del modelo retocado. Por último, se chequean los supuestos del modelo final y, en caso necesario, se realizan los ajustes pertinentes para evitar los problemas derivados del incumplimiento de los mismos.

Apéndice 1

Distribuciones de la familia exponencial

Fisher demostró en 1934 que la mayoría de las distribuciones de probabilidad que utilizamos al analizar datos son casos particulares de una amplia clase de distribuciones que el propio Fisher agrupó bajo la denominación de **familia exponencial**⁹. La teoría que da fundamento a la modelización lineal se basa en esta familia de distribuciones. Y en todos los modelos lineales que estudiaremos en este manual se asume que el componente aleatorio se ajusta a alguna distribución de la familia exponencial.

Consideremos la variable Y y su distribución¹⁰ de probabilidad $f(Y)$. Para destacar el hecho de que la distribución de probabilidad de Y depende de los parámetros θ y ϕ , la simbolizaremos mediante $f(Y|\theta, \phi)$ y la llamaremos distribución de probabilidad de la variable Y dados los parámetros θ y ϕ . Decimos que la función $f(Y|\theta, \phi)$ forma parte de la familia exponencial si puede expresarse de la siguiente manera:

$$f(Y|\theta, \phi) = \exp \left[\frac{Y\theta - b(\theta)}{a(\phi)} + c(Y, \phi) \right] \quad [1.17]$$

A θ se le llama parámetro **canónico** o **natural**; a ϕ , parámetro de **escala** o **dispersión**. Enseguida veremos que el primero tiene que ver con la media de Y y el segundo con su varianza.

La función propuesta en [1.17] está en forma **canónica**¹¹. El término $b(\theta)$ es el único que no depende de los datos. Además de ser un término clave para calcular los momentos de Y , su forma concreta determina el tipo de conexión (**enlace canónico**) que se establece entre la forma original de la función y la transformación que se aplica para obtener su forma canónica. En un modelo lineal, el enlace canónico se utiliza para hacer explícita la forma en que el componente sistemático se conecta con el componente aleatorio. Según veremos, tanto θ como $b(\theta)$ desempeñan un rol esencial en la modelización lineal.

La elección de las funciones b y c determina la forma concreta de [1.17]: normal, binomial, Poisson, gamma, etc. Y la media y la varianza de Y se obtienen a partir de las derivadas primera y segunda de $b(\theta)$ respecto de θ (ver, por ejemplo, Gill, 2001, págs. 24 y 27):

$$E(Y) = \frac{\partial[b(\theta)]}{\partial\theta} \quad \text{y} \quad V(Y) = a(\phi) \frac{\partial^2[b(\theta)]}{\partial\theta^2} \quad [1.18]$$

⁹ La mayoría de las distribuciones de probabilidad que se utilizan al analizar datos (normal, binomial, multinomial, Poisson, binomial negativa, gamma, ji-cuadrado, exponencial, gamma inversa, beta, pareto, etc.) forman parte de la familia exponencial. Hay, sin embargo, algunas distribuciones muy utilizadas que no forman parte de esta familia; por ejemplo, la distribución t de Student y la distribución uniforme.

¹⁰ El término *distribución de probabilidad* se refiere tanto a una *función de probabilidad discreta* como a una *función de densidad continua*.

¹¹ La forma canónica de una función es una simplificación práctica que se realiza con el objetivo de reducir la complejidad de la función para poder apreciar mejor su estructura y para facilitar el cálculo de los momentos. La transformación de una función a su forma canónica se realiza término a término, por lo que no se produce pérdida de información.

En este apartado se muestra cómo, efectivamente, algunas de las distribuciones más utilizadas pertenecen a la familia exponencial. Nos centraremos en las tres más utilizadas para representar el componente aleatorio de un modelo lineal: binomial, Poisson y normal.

La distribución binomial

La distribución *binomial* permite conocer la probabilidad asociada al *número de éxitos* en un conjunto de ensayos de Bernoulli, es decir, la probabilidad de obtener un determinado número de aciertos en un conjunto de respuestas, un determinado número de recuperaciones en un conjunto de pacientes tratados, etc. Por tanto, la distribución binomial sirve para trabajar con variables dicotómicas. Pero no exactamente con los dos valores que toma una variable dicotómica (uno-cero; éxito-fracaso), sino con el *número de éxitos* observados en un conjunto de n ensayos, registros o réplicas de una variable dicotómica. Siendo n_1 el *número de éxitos* y π_1 la *probabilidad de éxito*, las probabilidades binomiales asociadas a cada valor de n_1 se obtienen mediante

$$f(n_1 | n, \pi_1) = \binom{n}{n_1} \pi_1^{n_1} (1 - \pi_1)^{n - n_1} \quad [1.19]$$

Esta función ofrece las probabilidades asociadas a los diferentes valores de n_1 con el único requisito de que los ensayos sean independientes entre sí, es decir, con el único requisito de que π_1 permanezca constante en cada ensayo. Cuando se dan estas circunstancias (n ensayos de una variable dicotómica con probabilidad de éxito constante) decimos que la variable *número de éxitos* se distribuye binomialmente con parámetros¹² n y π_1 .

Aplicando unas sencillas transformaciones, la ecuación [1.19] puede expresarse en el formato de [1.17], es decir, en el formato de la familia exponencial:

$$f(n_1 | n, \pi_1) = \exp \left[\underbrace{n_1 \log_e \left(\frac{\pi_1}{1 - \pi_1} \right)}_{Y\theta} - \underbrace{(-n \log_e (1 - \pi_1))}_{b(\theta)} + \underbrace{\log_e \left(\binom{n}{n_1} \right)}_{c(Y)} \right] \quad [1.20]$$

(dado que el parámetro ϕ no aparece en la función, se asume $a(\phi) = 1$). Del primer término exponencial se deduce que el enlace canónico o parámetro natural de una distribución binomial es

$$\theta = \log_e [\pi_1 / (1 - \pi_1)] \quad [1.21]$$

Y con algo de álgebra, la función $b(\theta)$ puede expresarse en términos del parámetro natural θ como

$$b(\theta) = n \log_e [1 + \exp(\theta)] \quad [1.22]$$

Conociendo θ y $b(\theta)$ ya es posible aplicar las ecuaciones propuestas en [1.18] para obtener el valor esperado de $Y = n_1$ y su varianza (ver, por ejemplo, Gill, 2001, págs. 24 y 27):

$$E(Y) = E(n_1) = n\pi_1 \quad \text{y} \quad V(Y) = V(n_1) = n\pi_1(1 - \pi_1) \quad [1.23]$$

¹² Puesto que n y π_1 pueden tomar distintos valores, en realidad no existe una única distribución binomial, sino toda una familia de distribuciones binomiales (tantas como valores distintos puedan tomar n y π_1), todas las cuales se ajustan a la misma regla.

Por tanto, en una distribución binomial, el tamaño de la varianza es proporcional al tamaño de la media (el tamaño de la media determina el tamaño de la varianza). Esto es algo que será necesario tener muy en cuenta en la modelización lineal.

Según veremos, la distribución binomial se utiliza para modelar las respuestas dicotómicas del análisis de regresión logística.

La distribución de Poisson

La distribución de Poisson se utiliza para modelar *frecuencias* (el número de veces que se repite cada patrón de variabilidad) y *recuentos* (el número de ocurrencias de un determinado evento en un determinado intervalo de tiempo). La distribución de Poisson asume que, para intervalos cortos de tiempo, la probabilidad de que ocurra un determinado evento es fija y proporcional a la longitud del intervalo. Esta probabilidad se obtiene a partir de un único parámetro que es a la vez la media y la varianza de la distribución.

Llamando Y al *número de eventos* y μ al *número esperado de eventos*, las probabilidades que ofrece la distribución de Poisson para cada valor de Y vienen dadas por

$$f(Y|\mu) = \mu^Y e^{-\mu} / (Y!) \quad [1.24]$$

Unas sencillas transformaciones permiten expresar esta ecuación en el formato de la familia exponencial

$$f(Y|\mu) = \exp \left[\underbrace{Y \log_e(\mu)}_{Y\theta} - \underbrace{\mu}_{b(\theta)} + \underbrace{\log_e(Y!)}_{c(Y)} \right] \quad [1.25]$$

(de nuevo, puesto que ϕ no aparece en la ecuación, se verifica $a(\phi) = 1$). De la correspondencia establecida entre los términos exponenciales de [1.25] y los de [1.17] se sigue

$$\begin{aligned} \theta &= \log_e(\mu) \\ b(\theta) &= \mu = \exp(\theta) \end{aligned} \quad [1.26]$$

Y aplicando [1.18] se obtiene (ver, por ejemplo, Gill, 2001, págs. 23 y 26)

$$E(Y) = \mu \quad \text{y} \quad V(Y) = \mu \quad [1.27]$$

Por tanto, la distribución de Poisson tiene un único parámetro que es al mismo tiempo la media y la varianza de la distribución. Y, de nuevo, el hecho de que la varianza esté relacionada con la media es algo que será necesario tener muy en cuenta en la modelización lineal.

La distribución normal

La distribución normal es, sin duda, el referente más importante para un analista de datos. Es la elección habitual para modelar respuestas cuantitativas. Se trata de una distribución con dos parámetros (un parámetro de posición, μ , y un parámetro de escala, σ) cuya expresión habitual

$$f(Y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{- (Y - \mu)^2 / (2\sigma^2)} \quad [1.28]$$

puede transformarse fácilmente al formato de la familia exponencial:

$$f(Y|\mu, \sigma) = \exp \left[\underbrace{\left(Y\mu - \frac{\mu^2}{2} \right)}_{Y\theta} / \underbrace{\sigma^2}_{b(\theta)} + \underbrace{\frac{-1}{2} \left(\frac{Y^2}{\sigma^2} + \log_e(2\pi\sigma^2) \right)}_{a(\varphi)} \right] \quad [1.29]$$

De la correspondencia establecida entre los términos exponenciales de [1.29] y los de [1.17] se sigue

$$\begin{aligned} \theta &= \mu \\ a(\varphi) &= \sigma^2 \\ b(\theta) &= \mu^2/2 = \theta^2/2 \end{aligned} \quad [1.30]$$

Y aplicando [1.18] se obtiene (ver, por ejemplo, Gill, 2001, págs. 24 y 27)

$$E(Y) = \mu \quad \text{y} \quad V(Y) = \sigma^2 \quad [1.31]$$

A diferencia de lo que ocurre en las distribuciones binomial y Poisson, el tamaño de la varianza de una distribución normal es independiente del tamaño de la media. De hecho, la distribución normal es la única distribución de la familia exponencial en la que la varianza es independiente de la media.

Máxima verosimilitud

Los parámetros de un modelo lineal son valores desconocidos que es necesario estimar para que el modelo tenga alguna utilidad. Existen diferentes estrategias para efectuar estas estimaciones, pero la más utilizada cuando se trabaja con modelos mixtos y generalizados se conoce como método de *máxima verosimilitud*. Para una aproximación intuitiva a este método de estimación puede consultarse el Apéndice 7 del primer volumen (Pardo, Ruiz y San Martín, 2009); en Amón (1984, págs. 249-254) puede encontrarse una explicación algo más formal, muy clara y asequible incluso si se carece de una buena base matemática; y si se está dispuesto a profundizar algo más en todo lo relativo a la estimación por máxima verosimilitud puede consultarse Dunteman y Ho (2006, págs. 23-31), Harrell (2001, págs. 179-213) o Gill (2001, págs. 39-51).

La función de verosimilitud

Consideremos la variable aleatoria Y_i y su función de probabilidad $f(Y)$, y llamemos θ al parámetro (o conjunto de parámetros) involucrados en $f(Y)$. En un escenario de estas características, las probabilidades $f(Y)$ dependen tanto del valor de θ como de los valores concretos que tome Y_i . Como consecuencia de esto, $f(Y)$ puede interpretarse de dos maneras distintas. En primer lugar, como una **función de probabilidad** (o de densidad de probabilidad), en cuyo caso se considera que las probabilidades $f(Y)$ dependen del parámetro θ , el cual se asume conocido. Para enfatizar que θ es conocido y que las probabilidades de Y_i dependen de θ , la función de probabilidad de Y_i se simboliza mediante $f(Y|\theta)$. En segundo lugar, $f(Y)$ puede interpretarse como una **función de verosimilitud**, en cuyo caso se considera que la variable Y_i representa un conjunto de datos conocidos y θ es un parámetro (o conjunto de parámetros) desconocido cuyo

valor depende de Y_i . Para enfatizar que el valor de θ es desconocido y que depende de los valores concretos de Y_i , la función de verosimilitud se simboliza mediante $L(\theta|Y)$.

Desde un punto de vista estrictamente matemático, una función de probabilidad y una función de verosimilitud son la misma cosa¹³, es decir, $f(Y|\theta) = L(\theta|Y)$. Pero la primera interpreta los parámetros como fijos y los datos como variables y la segunda interpreta los datos como fijos y los parámetros como variables. Distinguir entre ambas funciones y utilizar cada una en su contexto suele facilitar las cosas.

Aclaremos el concepto de función de verosimilitud con un ejemplo concreto. Consideremos una variable categórica X_i con I categorías y llamemos $\mathbf{n}_i = (n_1, n_2, \dots, n_i, \dots, n_I)$ a las frecuencias obtenidas al seleccionar una muestra aleatoria de n casos y clasificarlos en las I categorías de X_i . Asumiendo que las probabilidades concretas del resultado muestral obtenido vienen dadas por la distribución *multinomial* con parámetros $\pi_1, \pi_2, \dots, \pi_i, \dots, \pi_I$, la *función de verosimilitud* de $\mathbf{n}_i$ es la función de probabilidad conjunta del resultado muestral $\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_i, \dots, \mathbf{n}_I$ dados los parámetros $\pi_1, \pi_2, \dots, \pi_i, \dots, \pi_I$, es decir,

$$L(\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_i, \dots, \mathbf{n}_I; \pi_1, \pi_2, \dots, \pi_i, \dots, \pi_I)$$

o, abreviadamente:

$$L(\mathbf{n}_i; \pi_i) \quad \text{con } i = 1, 2, \dots, I \quad [1.32]$$

Por tanto, una función de verosimilitud es una función que asigna probabilidades concretas a los valores muestrales obtenidos (de modo similar a como lo hace una función de probabilidad). Esas probabilidades dependen, en primer lugar, de la distribución de probabilidad elegida, que es una *distribución conocida* que se elige en función de las características de los datos; y, en segundo lugar, de los *parámetros desconocidos* de esa distribución de probabilidad, que justamente por ser desconocidos es por lo que necesitan ser estimados.

Estimación por máxima verosimilitud

La función de verosimilitud $L(\theta|Y)$ es la base de un método de estimación ampliamente utilizado llamado **máxima verosimilitud** (Fisher, 1925). La lógica de este método es bastante simple: consiste en utilizar como estimadores de los parámetros desconocidos θ los valores $\hat{\theta}$ que maximizan la probabilidad de obtener los datos de hecho obtenidos. Este método interpreta $L(\theta|Y)$ como función de θ y asigna a θ los valores $\hat{\theta}$ que maximizan¹⁴ $L(\theta|Y)$. El máximo de la función $L(\theta|Y)$ puede encontrarse igualando a cero su derivada parcial respecto de θ .

La estimación por máxima verosimilitud comienza con un conjunto de datos (por ejemplo, el resultado muestral $\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_i, \dots, \mathbf{n}_I$) y un modelo teórico de probabilidad (por ejemplo, la distribución *multinomial*) que expresa la probabilidad de ese conjunto de datos como función de uno o más parámetros desconocidos (por ejemplo, $\pi_1, \pi_2, \dots, \pi_i, \dots, \pi_I$). El objetivo de la estimación es encontrar los valores de esos parámetros desconocidos que hacen más probables (más verosímiles) los datos obtenidos: *las estimaciones de máxima verosimilitud son los valores que*

¹³ Aunque conviene señalar que las verosimilitudes no son exactamente probabilidades, pues no tienen todas sus propiedades. Entre otras cosas, no siempre las verosimilitudes de una variable categórica suman 1; ni tampoco se obtiene siempre 1 al integrar las verosimilitudes de una variable cuantitativa (ver por ejemplo, Ríos, 1977, pág. 328).

¹⁴ Con la única condición de que los valores de $\hat{\theta}$ se encuentren dentro del rango de valores asumibles por θ (por ejemplo, si θ es una varianza, $\hat{\theta}$ debe tomar un valor no negativo).

maximizan la función de verosimilitud. Por ejemplo, las estimaciones de máxima verosimilitud de los parámetros $\pi_1, \pi_2, \dots, \pi_i, \dots, \pi_I$ son los valores $\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_i, \dots, \hat{\pi}_I$ que hacen más probable el resultado muestral $n_1, n_2, \dots, n_i, \dots, n_I$. El máximo de $L(\theta|Y)$ puede encontrarse igualando a cero su derivada parcial respecto de θ .

Veamos cómo hacer esto con un ejemplo concreto. Consideremos una variable categórica X_i con I categorías y llamemos $n_i = (n_1, n_2, \dots, n_i, \dots, n_I)$ a las frecuencias obtenidas al seleccionar una muestra aleatoria de n casos y clasificarlos en las I categorías de X_i . Asumiendo que las probabilidades concretas del resultado muestral obtenido vienen dadas por la distribución de probabilidad *multinomial* con parámetros $\pi_1, \pi_2, \dots, \pi_i, \dots, \pi_I$, la frecuencia de una casilla concreta, pongamos n_1 , es una variable aleatoria cuya distribución de probabilidad es la binomial¹⁵ con parámetros n y π_1 . Por tanto,

$$L(n_1; \pi_1) = \pi_1^{n_1} (1 - \pi_1)^{n - n_1} \quad [1.33]$$

La estimación máximo-verosímil del parámetro π_1 se obtiene maximizando la función de verosimilitud L . Ahora bien, maximizar L equivale a maximizar su logaritmo (lo llamaremos LL), que al ser una función monótona no altera el máximo de la función original, y posee la propiedad de simplificar el proceso:

$$LL = \log_e(L) = n_1 \log_e(\pi_1) + (n - n_1) \log_e(1 - \pi_1) \quad [1.34]$$

Para encontrar el máximo de la función LL respecto al parámetro π_1 , basta con derivarla respecto a π_1 , igualar el resultado a cero y resolver:

$$\frac{\partial LL}{\partial \pi_1} = \frac{n_1}{\pi_1} - \frac{n - n_1}{1 - \pi_1} = 0 \quad [1.35]$$

de donde:

$$\begin{aligned} (1 - \pi_1)n_1 &= \pi_1(n - n_1) \\ n_1 &= n\pi_1 \\ \pi_1 &= n_1/n \end{aligned} \quad [1.36]$$

En consecuencia, $\hat{\pi}_1 = n_1/n$ es la estimación que maximiza la función [1.34]. Lo cual significa que las proporciones muestrales son las estimaciones máximo-verosímiles de las proporciones poblacionales. Como, por otro lado, la estimación obtenida será la misma cualquiera que sea la muestra utilizada, podemos decir de $\hat{\pi}_1$ que es el estimador de máxima verosimilitud de π_1 .

Por supuesto, si en lugar de estimar π_1 se desea estimar π_2 se obtendrá un resultado equivalente: $\hat{\pi}_2 = n_2/n$. Y lo mismo ocurrirá con cualquier otro parámetro π_j . Por tanto, el resultado

¹⁵ Cada una de las n observaciones puede considerarse un ensayo de Bernoulli con dos resultados posibles: cada observación puede ser clasificada en la primera categoría de X_i con probabilidad π_1 y en cualquier otra categoría distinta de la primera con probabilidad $1 - \pi_1$. En consecuencia, la función de probabilidad conjunta de las n observaciones o variables aleatorias $X_1, X_2, \dots, X_i, \dots, X_n$ (con $X_i = 1$ si una observación es clasificada en la primera categoría de X_i y $X_i = 0$ si una observación no es clasificada en la primera categoría de X_i) vendrá dada por:

$$\begin{aligned} L(X_1, X_2, \dots, X_n; \pi_1) &= L(n_1; \pi_1) \\ &= [\pi_1^{X_1} (1 - \pi_1)^{1 - X_1}] [\pi_1^{X_2} (1 - \pi_1)^{1 - X_2}] \dots [\pi_1^{X_n} (1 - \pi_1)^{1 - X_n}] \\ &= \pi_1^{n_1} (1 - \pi_1)^{n - n_1} \end{aligned}$$

obtenido puede generalizarse afirmando que las proporciones $\hat{\pi}_i = n_i/n$ son los estimadores de máxima verosimilitud de π_i .

Al margen de su simplicidad conceptual y algebraica, los estimadores de máxima verosimilitud poseen algunas importantes propiedades. En primer lugar, son estimadores *consistentes*: ofrecen estimaciones muy próximas al verdadero valor del parámetro y convergen con él conforme aumenta el tamaño muestral. En segundo lugar, cuando se trabaja con distribuciones de la familia exponencial, son estimadores *suficientes*: extraen de los datos toda la información necesaria para efectuar las estimaciones. Finalmente, conforme el tamaño muestral va aumentando, la distribución muestral de los estimadores máximo-verosímiles se va aproximando a la distribución normal.

En algunos casos, como el que acabamos de proponer sobre un conjunto de frecuencias distribuidas binomialmente, las estimaciones de máxima verosimilitud pueden derivarse analíticamente resolviendo las correspondientes derivadas parciales. En casos más complejos, la solución no es tratable analíticamente y es necesario aplicar algoritmos de cálculo iterativo. Nelder y Wedderburn (1972) han propuesto una técnica de cálculo llamada *mínimos cuadrados ponderados iterativamente* que permite obtener las estimaciones de máxima verosimilitud de cualquier modelo lineal que utilice para representar el componente aleatorio una distribución de la familia exponencial. Esta técnica de cálculo se basa en el método de tanteo *-scoring-* de Fisher y en el algoritmo de Newton-Raphson (de ambos puede encontrarse una buena descripción en Gill, 2001, págs. 39-51).

2

Modelos lineales clásicos

La expresión **modelos lineales clásicos** incluye, básicamente, los modelos de regresión lineal y los modelos de análisis de varianza y covarianza. Todos ellos son concreciones de lo que en estadística se conoce como **modelo lineal general** (no confundir *general* con *generalizado*) y a todos ellos prestaremos atención aquí, aunque el análisis de varianza y el análisis de regresión lineal ya hemos empezado a estudiarlos en el segundo volumen (ver Capítulos 6 al 10).

Lo que tienen en común estos modelos es que han sido diseñados para modelar **respuestas cuantitativas**. Todos ellos utilizan una función de enlace identidad y asumen que la varianza del componente aleatorio es constante para cada patrón de variabilidad (lo cual es equivalente a asumir que el componente aleatorio se distribuye normalmente, pues entre las distribuciones de la familia exponencial, que son las que se utilizan en la modelización lineal, únicamente en la distribución normal se verifica que la varianza es independiente de la media).

En lo que se diferencian unos modelos clásicos de otros es en las características del predictor lineal, es decir, en las variables independientes que incluyen; en concreto, en la naturaleza de las mismas (categóricas o cuantitativas) y en el rol que desempeñan dentro del modelo. Mientras que el predictor lineal de un **análisis de varianza** solo incluye variables categóricas, el predictor lineal de un **análisis de regresión lineal** puede incluir tanto variables categóricas como cuantitativas; y en ambos casos interesa analizar el efecto de todas ellas. El predictor lineal de un **análisis de covarianza** también puede incluir variables categóricas y cuantitativas, pero las variables cuantitativas se incluyen en el modelo para controlar su efecto, no para analizarlo.

Si bien los tres modelos clásicos mencionados pertenecen a la misma familia y permiten cubrir objetivos casi idénticos (todos ellos sirven para estudiar la relación entre

una variable dependiente cuantitativa y una o más variables independientes), cada uno de ellos pone el énfasis en diferentes aspectos del análisis. El análisis de varianza se centra en comparar en la variable cuantitativa los grupos definidos por una o más variables categóricas; el análisis de covarianza hace lo mismo pero controlando el efecto de terceras variables; el análisis de regresión pone el énfasis en la predicción y en la identificación de las variables independientes que ayudan a entender o explicar el comportamiento de la variable dependiente.

Para aplicar un análisis de varianza o un análisis de regresión basta con lo ya estudiado en el segundo volumen. No obstante, en este capítulo haremos un breve repaso de ambas herramientas desde la perspectiva integradora de la modelización lineal. Y empezaremos a estudiar el análisis de covarianza.

Análisis de varianza

Los modelos de análisis de varianza (ANOVA) más utilizados en el ámbito de las ciencias sociales y de la salud ya los hemos estudiado en el Volumen 2 (ver Pardo y San Martín, 2010, Capítulos 6 al 9). Pero los hemos estudiado desde la perspectiva clásica, es decir, identificando, aislando y comparando las diferentes fuentes de variabilidad presentes en el diseño. Para empezar a familiarizarnos con el ajuste de modelos lineales, este apartado incluye una breve descripción del *análisis de varianza de un factor* desde la perspectiva de la modelización lineal¹.

Recordemos que para ajustar un modelo lineal hay que llevar a cabo cuatro tareas: (1) seleccionar el modelo, (2) estimar los parámetros que incluye y obtener los pronósticos, (3) evaluar la calidad del modelo y (4) chequear los supuestos.

Para ilustrar las diferentes partes del análisis utilizaremos el mismo ejemplo que en el Capítulo 6 del segundo volumen, es decir, el ejemplo sobre la relación entre el *nivel de ansiedad* (variable independiente o factor) y el *rendimiento académico* (variable dependiente o respuesta). Los datos se encuentran en el archivo *Ansiedad rendimiento*, el cual puede descargarse de la página web del manual.

Seleccionar el modelo

El análisis de varianza (ANOVA) no es un único modelo lineal sino toda una familia de modelos. Cada uno de estos modelos incorpora los elementos necesarios para describir una situación concreta, pero todos ellos asumen que el componente aleatorio se distribuye normalmente y todos ellos utilizan una función de enlace identidad.

Analizar los datos correspondientes a un diseño de **un factor** (una variable independiente categórica que define grupos y una variable dependiente cuantitativa en la cual se desea comparar los grupos) requiere formular dos modelos alternativos, uno indican-

¹ Para profundizar en los contenidos que se exponen en este apartado puede consultarse Maxwell y Delaney (2004, págs. 69-97).

do que *no existe efecto* del factor (modelo 0) y otro indicando que *sí existe efecto* del factor (modelo 1):

$$\text{Modelo 0: } Y_{ij} = \mu + E_{ij(0)} \quad [2.1]$$

$$\text{Modelo 1: } Y_{ij} = \mu + \alpha_j + E_{ij(1)} \quad [2.2]$$

(el subíndice i se refiere a los sujetos: $i = 1, 2, \dots, n$; el subíndice j , a los grupos o niveles del factor: $j = 1, 2, \dots, J$; puesto que en el modelo 0 no hay grupos, el subíndice j toma el mismo valor en todos los sujetos; los subíndices 0 y 1 indican de qué modelo se trata). Los términos que incluyen estos dos modelos se corresponden con los ya propuestos en la Figura 1.2 del capítulo anterior como partes integrantes de un modelo lineal. El significado de estos términos ya se ha explicado en el capítulo anterior a propósito de las ecuaciones [1.4] y [1.5], y en el apartado *Elementos de un modelo lineal clásico* del Apéndice 2 se explica con más detalle el significado de cada término. Es importante reparar en el hecho de que estos dos modelos únicamente difieren en el término α_j y que ese término es justamente el que representa el efecto del factor.

Puesto que los errores no intervienen en los pronósticos (recordemos que los errores no forman parte del predictor lineal y que su valor esperado es cero), el modelo 0 asigna el mismo pronóstico o valor esperado $\mu_{ij(0)}$ a todos los sujetos:

$$\mu_{ij(0)} = \mu \quad [2.3]$$

Por tanto, el modelo 0 afirma que todas las puntuaciones Y son iguales. En este escenario de ausencia de efecto del factor, los errores, es decir, las diferencias entre los valores observados y los esperados o pronosticados, son las desviaciones de cada valor Y respecto de su media total:

$$E_{ij(0)} = Y_{ij} - \mu_{ij(0)} = Y_{ij} - \mu \quad [2.4]$$

El modelo 1 hace algo distinto: asigna el mismo valor a los sujetos del mismo grupo, pero un valor distinto a cada grupo:

$$\mu_{ij(1)} = \mu_j \quad (\text{con } \mu_j = \mu + \alpha_j) \quad [2.5]$$

Por tanto, en el modelo 1 se está afirmando que las medias poblacionales difieren en la parte atribuible al efecto del factor. En este escenario, los errores son las desviaciones de cada valor Y respecto de la media de su grupo:

$$E_{ij(1)} = Y_{ij} - \mu_{ij(1)} = Y_{ij} - \mu_j \quad [2.6]$$

Los modelos 0 y 1 (ecuaciones [2.1] y [2.2]) se corresponden con las hipótesis que se ponen a prueba en el análisis de varianza de un factor. El **modelo 0** se corresponde con la **hipótesis nula**, es decir, con la hipótesis de que todas las medias poblacionales son iguales:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_J = \dots = \mu_J \quad (Y_{ij} = \mu + E_{ij(0)}) \quad [2.7]$$

El **modelo 1** se corresponde con la **hipótesis alternativa**, es decir, con la hipótesis de que no todas las medias poblacionales son iguales (o, lo que es lo mismo, con la hipótesis de que la variable independiente está relacionada con la dependiente):

$$H_1: \mu_j \neq \mu_{j'} \quad \text{para algún } j \text{ o } j' \quad (Y_{ij} = \mu + \alpha_j + E_{ij(1)}) \quad [2.8]$$

Retomando nuestro ejemplo sobre la relación entre el nivel de ansiedad y el rendimiento académico, el modelo 0 (que se corresponde con la hipótesis nula) afirma que el rendimiento medio es el mismo en los tres niveles de ansiedad, es decir, que el rendimiento es independiente del nivel de ansiedad. Y el modelo 1 (que se corresponde con la hipótesis alternativa) afirma que el rendimiento medio no es el mismo en los tres niveles de ansiedad, es decir, que el rendimiento está relacionado con el nivel de ansiedad.

Estimar los parámetros y obtener los pronósticos

La cuestión que interesa resolver al plantear los modelos 0 y 1, es decir, la razón por la cual se formulan estos dos modelos para analizar los datos de un diseño de un factor completamente aleatorizado es para averiguar si el *parámetro extra* que incluye el modelo propuesto (el modelo 1) permite mejorar el ajuste del modelo nulo (el modelo 0); dicho de otra forma, para averiguar si el modelo propuesto permite explicar el comportamiento de los datos mejor que el modelo nulo. La respuesta a esta cuestión pasa por comparar la calidad del ajuste de ambos modelos. Y ya sabemos que la calidad de un modelo lineal se evalúa comparando los valores observados y los pronosticados. Pero para obtener los pronósticos es necesario estimar los parámetros.

Para obtener estas estimaciones pueden utilizarse diferentes métodos, pero en los modelos de análisis de varianza suele utilizarse el método de **mínimos cuadrados**. Este método utiliza como estimadores de los parámetros los valores que minimizan la suma de las diferencias al cuadrado entre los valores observados y los pronosticados.

Dado que el estimador mínimo-cuadrático de una media poblacional es su correspondiente media muestral (ver el apartado *Estimación por mínimos cuadrados*, en el Apéndice 7 del primer volumen), tenemos:

$$\begin{aligned} \hat{\mu} &= \bar{Y} \\ \hat{\mu}_j &= \bar{Y}_j \\ \hat{\alpha}_j &= \hat{\mu}_j - \hat{\mu} = \bar{Y}_j - \bar{Y} \end{aligned} \quad [2.9]$$

Una vez estimados los parámetros, es posible obtener los pronósticos que se derivan de ambos modelos simplemente sustituyendo los parámetros de [2.3] y [2.5] por sus correspondientes estimadores:

$$\text{Pronósticos del modelo 0: } \hat{\mu}_{ij(0)} = \hat{\mu} = \bar{Y} \quad [2.10]$$

$$\text{Pronósticos del modelo 1: } \hat{\mu}_{ij(1)} = \hat{\mu}_j = \bar{Y}_j \quad [2.11]$$

En nuestro ejemplo sobre la relación entre el nivel de ansiedad y el rendimiento académico, el modelo 0 pronostica el mismo valor a los 30 participantes en el estudio: el rendimiento medio ($\bar{Y}_{\text{total}} = 10$; las medias observadas pueden consultarse en la Tabla 6.2 del segundo volumen). Mientras que el modelo 1 ofrece tres pronósticos distintos, uno para cada grupo o nivel de ansiedad ($\bar{Y}_{\text{bajo}} = 9$, $\bar{Y}_{\text{medio}} = 14$, $\bar{Y}_{\text{alto}} = 7$).

Valorar la calidad o ajuste del modelo

El paralelismo existente entre el enfoque basado en el contraste de hipótesis clásico y el enfoque basado en la modelización lineal es claro: el punto de partida en el primero es la formulación de dos hipótesis que representan la ausencia y la presencia del efecto estudiado; el punto de partida en el segundo es la formulación de dos modelos alternativos que representan, también, la ausencia y la presencia del efecto estudiado. En el enfoque clásico, el rechazo de la hipótesis nula permite concluir que el efecto estudiado es no nulo; en la modelización lineal, la comparación entre ambos modelos permite determinar si el término en el que difieren es no nulo.

La comparación entre modelos se basa en un estadístico llamado **desvianza** ($-2LL$). Este estadístico se obtiene a partir de las discrepancias existentes entre los valores *observados* (Y_{ij}) y los *pronosticados* ($\hat{\mu}_{ij}$). Los valores pronosticados se obtienen a partir de [2.10] y [2.11]. Y las discrepancias entre los valores observados y los pronosticados se obtienen estimando los errores definidos en [2.4] y [2.6], es decir, sustituyendo los parámetros μ y μ_j de esas ecuaciones por sus respectivos estimadores. A estas versiones muestrales de los errores poblacionales las llamamos **residuos** y los representamos mediante $\hat{E}$:

$$\text{Residuos del modelo 0: } \hat{E}_{ij(0)} = Y_{ij} - \hat{\mu} = Y_{ij} - \bar{Y} \quad [2.12]$$

$$\text{Residuos del modelo 1: } \hat{E}_{ij(1)} = Y_{ij} - \hat{\mu}_j = Y_{ij} - \bar{Y}_j \quad [2.13]$$

El estadístico $-2LL$ se basa en estos residuos, concretamente en la suma de sus cuadrados:

$$-2LL_0 = \sum_i \sum_j \hat{E}_{ij(0)}^2 = \sum_i \sum_j (Y_i - \bar{Y})^2 \quad [2.14]$$

$$-2LL_1 = \sum_i \sum_j \hat{E}_{ij(1)}^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_j)^2 \quad [2.15]$$

Estas ecuaciones permiten apreciar que la desvianza ($-2LL$) es tanto mayor cuanto mayor es la diferencia entre los valores observados y los pronosticados. Por tanto, la desvianza refleja el grado de *desajuste* de un modelo, es decir, el grado en que un modelo se aleja del ajuste perfecto. La desvianza del modelo 0 ($-2LL_0$) refleja el máximo desajuste posible (el desajuste que resulta al pronosticar la variable dependiente sin otra información que la propia variable dependiente). La desvianza del modelo 1 ($-2LL_1$) refleja el desajuste del modelo que incorpora la información de la variable indepen-

diente. La diferencia entre ambas desvianzas refleja la diferencia en el desajuste de ambos modelos. Dividiendo esa diferencia entre la desviación del modelo 0 se obtiene R^2 (un estadístico que ya conocemos y que expresa la proporción en que el modelo 1 consigue reducir el desajuste del modelo 0). Y dividiendo esa diferencia entre la desviación del modelo 1 se obtiene una estimación de la *proporción en que aumenta el desajuste (PAD)* al eliminar del modelo 1 la variable independiente o factor:

$$PAD = \frac{-2LL_0 - (-2LL_1)}{-2LL_1} \quad [2.16]$$

Cuanto mayor es el valor de este estadístico, mayor es la diferencia en el desajuste de ambos modelos. Si la variable independiente o factor realmente está contribuyendo a reducir el desajuste, entonces el desajuste del modelo 1 debe ser significativamente menor que el del modelo 0.

Dividiendo el numerador y el denominador de [2.16] entre sus respectivos grados de libertad (gl), se obtiene un estadístico que permite valorar la diferencia en el desajuste de ambos modelos.

$$F = \frac{[-2LL_0 - (-2LL_1)] / (gl_0 - gl_1)}{-2LL_1 / gl_1} \quad [2.17]$$

Bajo ciertas condiciones (ver siguiente apartado) este estadístico se aproxima a la distribución F con $J - 1$ y $N - J$ grados de libertad. Por tanto, se trata de un estadístico que contiene la información necesaria y suficiente para contrastar la hipótesis nula de que el término extra que incluye el modelo 1 vale cero en la población (es decir, la hipótesis nula de que $\alpha_j = 0$ para todo j). O, lo que es equivalente, que las J medias poblacionales son iguales³. Un estadístico F significativo ($p < 0,05$) permitirá rechazar esta hipótesis nula y concluir que los J promedios poblacionales comparados no son iguales⁴. O, lo que es equivalente, permitirá concluir que la variable independiente o factor está relacionada con la variable dependiente.

² Los grados de libertad del modelo 0 son $N - 1$: puesto que en este modelo únicamente se está estimando un parámetro (la media total), solo se pierde un grado de libertad. Los grados de libertad del modelo 1 son $N - J$: puesto que en este modelo se están estimando J parámetros (las medias de los J grupos), se pierden J grados de libertad. Por tanto, los grados de libertad del numerador son $(N - 1) - (N - J) = J - 1$ (N se refiere al número total de casos y J al número de grupos).

³ Unas sencillas transformaciones permiten comprobar que el estadístico F propuesto en [2.17] es exactamente el mismo que ya hemos utilizado en el Capítulo 6 del segundo volumen, ecuación [6.6], para contrastar la hipótesis nula de igualdad de medias:

$$F = \frac{[\sum_i \sum_j (x_{ij} - \bar{Y})^2 - \sum_i \sum_j (x_{ij} - \bar{Y}_j)^2] / [(N - 1) - (N - J)]}{\sum_i \sum_j (x_{ij} - \bar{Y}_j)^2 / (N - J)} = \frac{\sum_j n_j (\bar{X}_j - \bar{Y})^2 / (J - 1)}{\sum_j (n_j - 1) S_j^2 / (N - J)} = \frac{MC_A}{MC_E}$$

⁴ Existen múltiples procedimientos para determinar qué media en concreto difiere de qué otra (ver, en el Capítulo 6 del segundo volumen, el apartado *Comparaciones múltiples entre medias*).

A las desviaciones de un ANOVA se les llama *sumas de cuadrados*. El SPSS las incluye en la tabla resumen del ANOVA (ver la Tabla 6.6 del segundo volumen): la desviación del modelo 0 ($-2LL_0$) es la *suma de cuadrados total*, la desviación del modelo 1 ($-2LL_1$) es la *suma de cuadrados intragrupos* o *error*. La diferencia entre ambas desviaciones es la *suma de cuadrados intergrupos*. En nuestro ejemplo sobre la relación entre el nivel de ansiedad y el rendimiento académico tenemos: $-2LL_0 = 614$ y $-2LL_1 = 354$. Colocando estos valores en [2.16] obtenemos

$$PAD = \frac{614 - 354}{354} = 0,73$$

Este resultado indica que el desajuste del modelo 0 (el modelo que únicamente incluye el término constante) es un 73% mayor que el del modelo 1 (el modelo que incluye el término constante y el efecto del nivel de ansiedad). Dividiendo la diferencia $614 - 354 = 260$ entre la desviación del modelo 0, es decir, entre 614, se obtiene $R^2 = 0,42$, valor que indica que el modelo 1 reduce el desajuste del modelo 0 en un 42%.

El estadístico F propuesto en [2.17] se obtiene dividiendo el numerador y el denominador de [2.16] entre sus respectivos grados de libertad: $J - 1 = 3 - 1 = 2$ para el numerador y $N - J = 30 - 3 = 27$ para el denominador. Por tanto,

$$F = \frac{(614 - 354)/2}{354/27} = \frac{130}{13,11} = 9,92$$

La probabilidad de encontrar valores mayores que 9,92 en la distribución F con 2 y 27 grados de libertad vale 0,0006 (esta probabilidad puede obtenerse en SPSS con la función CDF.F de la opción **Calcular**). Por tanto, al eliminar del modelo 1 el término correspondiente al efecto del factor se produce un aumento significativo del desajuste.

Chequear los supuestos

Para que un modelo lineal funcione correctamente es necesario que se den ciertas condiciones. Estas condiciones son las que garantizan que las estimaciones que realizamos son insesgadas y eficientes, y que el estadístico F funciona como se espera que lo haga.

En el caso del análisis de varianza de un factor estas condiciones son las que ya hemos llamado abreviadamente *independencia*, *normalidad* e *igualdad de varianzas* (en caso necesario, consultar el apartado *Supuestos del ANOVA de un factor*, en el Capítulo 6 del segundo volumen; lo dicho allí, sirve también aquí).

Cuando aplicamos un análisis de varianza de un factor asumimos que estamos trabajando con J muestras aleatorias procedentes de poblaciones normales con la misma varianza. También asumimos que lo que una puntuación se desvía del promedio de su población (E_{ij}) es *independiente* de lo que se desvía otra puntuación cualquiera de esa misma población: $COV(E_{ij}, E_{i'j'}) = 0$ (siendo i e i' dos casos distintos). Y también asumimos que las desviaciones (errores) de cada puntuación respecto de su media son aleatorias y unas se anulan con otras: $E(E_{ij}) = 0$.

Análisis de covarianza

En los modelos lineales propuestos hasta ahora hemos asumido que el efecto de terceras variables sobre la relación estudiada forma parte del conjunto de efectos no tenidos en cuenta. Ya nos hemos referido a estas variables como *concomitantes* o *extrañas* y, para neutralizar su efecto, hemos propuesto aplicar técnicas de control experimental como asignar aleatoriamente los sujetos a las condiciones del estudio, formar bloques aleatorios o mantener constantes las condiciones de aplicación de los tratamientos.

En este apartado vamos a estudiar una estrategia alternativa de control: el **análisis de covarianza** (ANCOVA). No se trata de una estrategia de control experimental, pues no se basa en la modificación de las condiciones del diseño, sino de control **estadístico**, pues, según veremos, se basa en la aplicación combinada del análisis de varianza y del análisis de regresión. Al igual que con el experimental, con el control estadístico se pretende reducir la variabilidad error y aumentar la precisión de las estimaciones.

Supongamos que se ha llevado a cabo un estudio para comparar la eficacia de dos tratamientos antidepresivos. Supongamos además que quienes han diseñado el estudio tienen la sospecha de que los pacientes más jóvenes podrían recuperarse mejor. Tenemos un diseño con una variable *independiente* (los tratamientos), una variable *dependiente* (la recuperación) y una variable *extraña* (la edad).

La forma habitual de comparar la eficacia de dos tratamientos consiste en valorar los resultados que se obtienen con cada uno al administrarlos a muestras aleatorias de pacientes. La asignación aleatoria es la mejor estrategia de que disponemos para intentar hacer que los grupos con los que vamos a trabajar sean *equivalentes*, es decir, la mejor forma que tenemos de controlar el conjunto de efectos no tenidos en cuenta.

Pero la equivalencia entre grupos que se consigue con la asignación aleatoria puede mejorarse aplicando algún tipo de control sobre el efecto de las variables sospechosas de estar alterando los resultados del estudio (la *edad* en nuestro ejemplo). Una forma de control muy utilizada consiste en asignar los sujetos a los tratamientos después de formar *bloques aleatorios* (grupos de pacientes con la misma o parecida edad). Pero ocurre que no siempre es posible formar bloques aleatorios. No es posible, por ejemplo, cuando la variable cuyo efecto se desea controlar no se conoce antes de asignar los sujetos a las condiciones del estudio; y, lo que es más habitual, tampoco es posible formar bloques aleatorios cuando se trabaja con grupos intactos como los alumnos de una clase, los pacientes de un hospital o los votantes de un distrito (situaciones, todas ellas, en las que no existe asignación aleatoria de los sujetos a las condiciones del estudio).

Si cada uno de los tratamientos de nuestro ejemplo se administrara a pacientes de un hospital distinto, lo que se estaría eligiendo aleatoriamente sería el hospital, no los pacientes. Al hacer esto, ni se estaría utilizando asignación aleatoria de los pacientes a las condiciones del estudio ni se estarían formando bloques aleatorios (las dos formas habituales de control de terceras variables). Sin embargo, en estos casos todavía sería posible controlar el efecto de terceras variables aplicando herramientas de control estadístico como el análisis de covarianza. Según veremos, controlar el efecto de terceras variables tiene dos beneficios claros: (1) disminuye la variabilidad error, lo cual hace

aumentar la potencia de los contrastes que se llevan a cabo, y (2) elimina ruido del modelo y, con ello, se reduce el sesgo de las estimaciones.

Lógica del análisis de covarianza

Al analizar los datos correspondientes a un diseño de **un factor** completamente aleatorizado hemos formulado el siguiente modelo de ANOVA (modelo 1):

$$Y_{ij} = \mu + \alpha_j + E_{ij} \quad [2.2, \text{repetida}]$$

Para incluir en este modelo el efecto de una *covariable* (variable cuyo efecto deseamos controlar) basta con añadir un término que represente la relación entre esa covariable y la variable dependiente. Esto suele hacerse de la siguiente manera:

$$Y_{ij} = \mu + \alpha_j + \beta x_{ij} + E'_{ij} \quad [2.18]$$

El parámetro β es el coeficiente de regresión de Y_{ij} sobre x_{ij} ; por tanto, representa el grado de relación lineal existente entre la covariable y la variable dependiente. La letra x minúscula indica que se trata de puntuaciones diferenciales o de desviación.

Si la covariable elegida fuera completamente independiente de la variable dependiente, el parámetro β valdría cero y el modelo propuesto en [2.18] sería idéntico al propuesto en [2.2], en cuyo caso no habríamos ganado nada incorporando al modelo el efecto de la covariable. Simplemente trasladando el término βx_{ij} a la parte izquierda de [2.18] obtenemos

$$Y_{ij} - \beta x_{ij} = \mu + \alpha_j + E'_{ij} \quad [2.19]$$

Y esta expresión permite apreciar que en un modelo de ANCOVA se está utilizando el mismo componente sistemático ($\mu + \alpha_j$) que en un modelo de ANOVA. La diferencia entre ambos modelos es que, mientras que en un ANOVA se está intentando explicar o describir la variable dependiente Y_{ij} , en un ANCOVA se está intentando explicar o describir esa misma variable tras eliminar de ella, mediante regresión lineal, el efecto atribuible a la covariable.

Si la covariable elegida está relacionada con la variable dependiente, el parámetro β será distinto de cero. Y puesto que los términos Y_{ij} , μ y α_j son idénticos en ANOVA y en ANCOVA, una sencilla comparación de los modelos [2.2] y [2.18] permite apreciar que lo que realmente se está haciendo con un ANCOVA es eliminar de la variabilidad error la parte atribuible al efecto de la covariable:

$$E'_{ij} = E_{ij} - \beta x_{ij} \quad [2.20]$$

Por tanto, la primera consecuencia de incluir en el modelo el efecto de la covariable es la reducción del error, es decir, la reducción del término que representa al conjunto de efectos no tenidos en cuenta. Esta reducción será tanto mayor cuanto mayor sea la relación entre la covariable y la variable dependiente.

Seleccionar el modelo

Es posible formular tantos modelos de ANCOVA como de ANOVA: con un factor, con más de un factor; con efectos fijos, con efectos aleatorios; completamente aleatorizados, con medidas repetidas; etc. Todos ellos utilizan una función de enlace identidad y asumen que el componente aleatorio se distribuye normalmente. La única diferencia entre los modelos de ANOVA y de ANCOVA está en los términos extra que incluyen los segundos para representar el efecto de las covariables (un término extra por cada covariable).

Para analizar un diseño de un factor mediante modelos de ANOVA hemos formulado dos modelos alternativos (ver ecuaciones [2.1] y [2.2]): el modelo 0, que no incluye el efecto del factor, y el modelo 1, que sí lo incluye; comparando ambos modelos es posible aislar y evaluar el efecto del factor. Al incorporar covariables aumenta el número de modelos que aportan información útil, pero los dos modelos que permiten aislar y evaluar el efecto del factor son los siguientes:

$$\text{Modelo 0: } Y_{ij} = \mu + \beta x_{ij} + E_{ij(0)} \quad [2.21]$$

$$\text{Modelo 1: } Y_{ij} = \mu + \alpha_j + \beta x_{ij} + E_{ij(1)} \quad [2.22]$$

El **modelo 0** es un modelo de regresión lineal de Y sobre X ; comparándolo con el modelo nulo del ANOVA de un factor (ver ecuación [2.1]) es posible aislar el término βx_{ij} y averiguar si la covariable está relacionada con la variable dependiente (esta relación no constituye el objetivo primordial del análisis pero, según veremos, tiene su importancia, pues da pistas acerca de la conveniencia o no de controlar el efecto de la covariable). El **modelo 1** es un modelo de ANCOVA; incluye tanto el efecto del factor (α_j) como el de la covariable (βx_{ij}); comparando este modelo con el modelo 0 es posible aislar el término α_j y valorar el efecto del factor tras controlar (tras eliminar de Y_{ij}) el efecto debido a la covariable; esta comparación es el objetivo principal de un ANCOVA.

La **hipótesis nula** que se contrasta con un modelo de ANCOVA es la misma que la que se contrasta con un modelo de ANOVA:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_j = \dots = \mu_J \quad [2.23]$$

El matiz que añade el modelo de ANCOVA es que la afirmación [2.23] se refiere a las medias *corregidas*, es decir, a las medias que se obtienen tras eliminar de la variable dependiente el efecto de la covariable. Por tanto, el modelo de ANCOVA que se corresponde con la hipótesis nula es el **modelo 0**. Para indicar que se trata de medias corregidas utilizaremos asteriscos:

$$H_0: \mu_1^* = \mu_2^* = \dots = \mu_J^* = \dots = \mu_J^* \quad (Y_{ij} = \mu + \beta x_{ij} + E_{ij(0)}) \quad [2.24]$$

La **hipótesis alternativa** afirma que las medias corregidas no son iguales. El modelo de ANCOVA que se corresponde con esta hipótesis es el **modelo 1**:

$$H_1: \mu_j^* \neq \mu_{j'}^* \quad \text{para algún } j \text{ o } j' \quad (Y_{ij} = \mu + \alpha_j + \beta x_{ij} + E_{ij(1)}) \quad [2.25]$$

Los pronósticos que se derivan de los modelos 0 y 1 son pronósticos corregidos por el efecto de la covariable:

$$\mu_{ij(0)} = \mu + \beta x_{ij} \quad [2.26]$$

$$\mu_{ij(1)} = \mu_j + \beta x_{ij} \quad (\text{con } \mu_j = \mu + \alpha_j) \quad [2.27]$$

Por tanto, al realizar pronósticos, tanto el modelo 0 como el 1 tienen en cuenta la relación existente entre la variable dependiente y la covariable (βx_{ij}). Pero los pronósticos del modelo 0 se basan en la *media total* y los del modelo 1 en la *media de cada grupo*. En este escenario, los errores de cada modelo vienen dados por

$$E_{ij(0)} = Y_{ij} - \mu_{ij(0)} = Y_{ij} - (\mu + \beta x_{ij}) \quad [2.28]$$

$$E_{ij(1)} = Y_{ij} - \mu_{ij(1)} = Y_{ij} - (\mu_j + \beta x_{ij}) \quad [2.29]$$

Estimar los parámetros y obtener los pronósticos

En los modelos de ANCOVA, al igual que en los de ANOVA, las estimaciones de los parámetros pueden obtenerse aplicando diferentes métodos. No obstante, lo habitual es obtener las estimaciones con el método de **mínimos cuadrados**, es decir, con el método que utiliza como estimadores de los parámetros los valores que minimizan las diferencias entre los valores observados y los esperados.

Los estimadores mínimo-cuadráticos de los parámetros μ , μ_j y α_j ya los conocemos (ver [2.9]). El estimador mínimo cuadrático del parámetro β cambia dependiendo del modelo. El modelo 0 define un escenario con una sola población; solo es necesario un parámetro β para describir la relación entre X e Y :

$$\hat{\beta}_{(0)} = S_{XY}/S_X^2 = \sum_i \sum_j (X_{ij} - \bar{X})(Y_{ij} - \bar{Y}) / \sum_i \sum_j (X_{ij} - \bar{X})^2 \quad [2.30]$$

El modelo 1 define un escenario con J poblaciones (una por cada nivel del factor). Para describir la relación entre X e Y hacen falta J parámetros β , uno para cada población. Pero el modelo 1 incluye un único parámetro β , no J . Ese único parámetro puede estimarse de diferentes maneras, pero la mejor de todas consiste en utilizar la media ponderada de las J estimaciones disponibles. Esto equivale a:

$$\hat{\beta}_{(1)} = \sum_i \sum_j (X_{ij} - \bar{X}_j)(Y_{ij} - \bar{Y}_j) / \sum_i \sum_j (X_{ij} - \bar{X}_j)^2 \quad [2.31]$$

Una vez obtenidos los estimadores de los parámetros, los pronósticos que se derivan de ambos modelos se obtienen simplemente sustituyendo los parámetros en [2.26] y [2.27] por sus correspondientes estimadores:

$$\text{Pronósticos del modelo 0: } \hat{\mu}_{ij(0)} = \hat{\mu} + \hat{\beta}_{(0)} x_{ij} \quad [2.32]$$

$$\text{Pronósticos del modelo 1: } \hat{\mu}_{ij(1)} = \hat{\mu}_j + \hat{\beta}_{(1)} x_{ij} \quad [2.33]$$

Valorar la calidad o ajuste del modelo

La forma de valorar el efecto del factor consiste en comparar el modelo 1 (que incluye el efecto del factor) con el modelo 0 (que no incluye el efecto del factor). Esta comparación entre modelos se basa en la **desvianza** ($-2LL$), la cual, recordemos, se obtiene a partir de las discrepancias existentes entre los valores *observados* (Y_{ij}) y los *pronosticados* ($\hat{\mu}_{ij}$). Los valores pronosticados se obtienen mediante [2.32] y [2.33]. Y las discrepancias entre los valores observados y los pronosticados, es decir, los residuos, se obtienen sustituyendo los parámetros de las ecuaciones [2.28] y [2.29] por sus correspondientes estimadores:

$$\text{Residuos del modelo 0: } \hat{E}_{ij(0)} = Y_{ij} - (\hat{\mu} + \hat{\beta}_{(0)} x_{ij}) \quad [2.34]$$

$$\text{Residuos del modelo 1: } \hat{E}_{ij(1)} = Y_{ij} - (\hat{\mu}_j + \hat{\beta}_{(1)} x_{ij}) \quad [2.35]$$

El estadístico $-2LL$ se basa en estos residuos, concretamente en la suma de sus cuadrados:

$$-2LL_0 = \sum_i \sum_j \hat{E}_{ij(0)}^2 = \sum_i \sum_j (Y_{ij} - \hat{\mu} - \hat{\beta}_{(0)} x_{ij})^2 \quad [2.36]$$

$$-2LL_1 = \sum_i \sum_j \hat{E}_{ij(1)}^2 = \sum_i \sum_j (Y_{ij} - \hat{\mu}_j - \hat{\beta}_{(1)} x_{ij})^2 \quad [2.37]$$

Estos estadísticos indican cuánto se aleja cada modelo del ajuste perfecto. Y siguiendo la lógica ya expuesta a propósito de los estadísticos [2.16] y [2.17], el cociente

$$F = \frac{[-2LL_0 - (-2LL_1)] / (gl_0 - gl_1)}{-2LL_1 / gl_1} \quad [2.38]$$

indica cuánto aumenta el desajuste al eliminar del modelo el efecto del factor. Bajo ciertas condiciones (ver siguiente apartado) este estadístico se aproxima a la distribución F con⁵ $J - 1$ y $N - J - 1$ grados de libertad. Y permite contrastar la hipótesis nula de que el término extra que incluye el modelo 1, es decir, α_j , vale cero para todo j ; o, lo que es lo mismo, la hipótesis nula de que, cuando todos los grupos puntúan igual en la covariable, las J medias poblacionales de Y son iguales.

Chequear los supuestos

Para que un modelo de ANOVA funcione correctamente debe darse una serie de condiciones. A estas condiciones las hemos llamado *supuestos* y, en el caso del modelo de

⁵ Los grados de libertad del modelo 0 son $N - 2$; se pierden 2 grados de libertad al estimar los parámetros μ y β . Los grados de libertad del modelo 1 son $N - J - 1$; se pierden $J + 1$ grados de libertad al estimar las J medias μ_j y el parámetro β . Por tanto, los grados de libertad del numerador de [2.38] son $(N - 2) - (N - J) = J - 1$ (N se refiere al número de casos y J al número de grupos).

un factor completamente aleatorizado, hemos mencionado tres a los que hemos llamado, abreviadamente, *independencia*, *normalidad* y *homocedasticidad*. Para que un modelo de ANCOVA funcione correctamente deben darse estas mismas tres condiciones más alguna adicional (relacionada con la presencia de la covariable) que exponemos a continuación⁶.

En primer lugar, en un modelo de ANCOVA se asume que la covariable es de efectos fijos y que su relación con la variable dependiente es lineal. Estas dos condiciones son idénticas a las ya estudiadas a propósito de la regresión lineal (en caso necesario, revisar el Capítulo 10 del segundo volumen). Por un lado, asumir que **los valores de la covariable son fijos**⁷ implica, por un lado, que la covariable está medida sin error (esto tiene su importancia, pues cuanto menos fiable es la medida, menos preciso es el contraste del efecto del factor; ver Maxwell y Delaney, 2004, págs. 427-428) y, por otro, que las inferencias que es posible hacer deben basarse en los valores concretos que toma la covariable, no en otros. Por otro lado, asumir que **la relación entre la covariable y la variable dependiente es lineal** es algo que viene impuesto por el propio modelo: para representar la relación entre la covariable y la variable dependiente se está utilizando una ecuación que estima para Y un cambio constante (lineal) de tamaño β por cada unidad que aumenta X ; y no tiene sentido utilizar una ecuación lineal si la relación subyacente no es lineal.

En segundo lugar, se asume que **el factor no afecta a la covariable**. Si se utiliza una covariable que puede verse afectada por la administración de los tratamientos, la covariable debe medirse antes de administrar los tratamientos (esto es lo que se hace, por ejemplo, con la medida *pre* en un diseño pre-post). Si se utiliza una covariable que no puede verse afectada por los tratamientos (como, por ejemplo, la edad) podrá registrarse tanto antes como después de administrar los tratamientos, pero habrá que vigilar que los grupos no tengan promedios muy distintos en ella, pues lo contrario podría llegar a complicar sensiblemente la interpretación de los resultados (hasta el punto de que algunos expertos sugieren no utilizar modelos de ANCOVA cuando se incumple este supuesto; ver Keppel y Wickens, 2004, págs. 337-341).

Por último, se asume que **las J pendientes de regresión** (una por cada nivel del factor) **son iguales**. Puesto que el modelo de ANCOVA incluye un solo parámetro β (es decir, una única pendiente para todos los casos y no una para cada grupo), se está asumiendo que el grado de relación lineal existente entre la covariable y la variable dependiente es el mismo en todos los grupos. Este supuesto tiene su importancia cuando se utiliza el modelo que incluye un único parámetro β para representar la relación entre X e Y en todos los grupos, pero, según veremos, existe la posibilidad de ajustar modelos que incluyen pendientes distintas para cada grupo.

⁶ El lector interesado en profundizar en los supuestos del ANCOVA algo más de lo que lo haremos aquí puede consultar Maxwell y Delaney (2004, págs. 420-428).

⁷ Aunque en la práctica esto no suele ser así (pensemos en un diseño pre-post en el que la medida *pre* se utiliza como covariable y la medida *post* como variable dependiente), si la fiabilidad de la medida es lo bastante buena (suelen considerarse aceptables valores mayores de 0,80), el contraste del efecto del factor no se verá afectado por el hecho de que la covariable sea de efectos fijos o de efectos aleatorios.

Análisis de covarianza con SPSS

La Tabla 2.1 recoge los datos obtenidos en un estudio diseñado para valorar el efecto de tres métodos de enriquecimiento motivacional (variable independiente o factor a la que llamaremos *método*) sobre el rendimiento académico (variable dependiente a la que llamaremos *rendimiento*). La tabla incluye una medida del cociente intelectual de los sujetos (covariable a la que llamaremos *CI*). La última fila informa del rendimiento medio de cada grupo y del CI medio de cada grupo. El rendimiento medio de todos los sujetos vale 5,93 y el CI medio 106. Los datos se encuentran en el archivo *Motivación rendimiento*, el cual puede descargarse de la página web del manual.

El SPSS incluye varios procedimientos para ajustar modelos de ANCOVA. En este apartado vamos a utilizar el procedimiento **Univariante** para ajustar el modelo de un factor de efectos fijos, completamente aleatorizado, a los datos de la Tabla 2.1 (en los dos próximos capítulos veremos cómo ajustar otros modelos de ANCOVA). El objetivo del análisis es valorar el efecto de los *métodos* sobre el *rendimiento* controlando el efecto del *cociente intelectual*.

Tabla 2.1. Método de entrenamiento, rendimiento académico (Y) y cociente intelectual (X)

<i>Método A</i>		<i>Método B</i>		<i>Método C</i>	
<i>Rendim. (Y_{i1})</i>	<i>CI (X_{i1})</i>	<i>Rendim. (Y_{i2})</i>	<i>CI (X_{i2})</i>	<i>Rendim. (Y_{i3})</i>	<i>CI (X_{i3})</i>
6	100	4	90	3	80
8	130	7	120	5	100
7	110	6	95	7	110
9	130	5	100	4	100
7	110	6	110	5	105
7,40	116	5,60	103	4,80	99

Cómo chequear los supuestos

Cuando se decide aplicar un modelo de ANCOVA es necesario, antes de cualquier otra consideración, detenerse a comprobar si se dan las condiciones para hacerlo. En lo relativo a la *independencia* de las observaciones y a la *normalidad* y *homocedasticidad* de las distribuciones vale lo ya dicho a propósito del ANOVA de un factor (ver Capítulo 6 del segundo volumen). Con los supuestos específicos del ANCOVA, es decir, con los supuestos que tienen que ver con la presencia de la covariable, hay que realizar tres comprobaciones:

1. *La covariable correlaciona linealmente con la variable dependiente.* Esto puede comprobarse con el coeficiente de correlación de Pearson o con un análisis de regresión lineal. En nuestro ejemplo, el coeficiente de correlación de Pearson entre

el CI y el rendimiento vale 0,91, con $p < 0,0005$; por tanto, parece que en la relación entre el CI y el rendimiento existe un componente lineal significativo.

2. *El factor no afecta a la covariable.* Esto puede chequearse mediante un ANOVA tomando la covariable como variable dependiente. Aplicando un ANOVA a los datos de nuestro ejemplo (con los métodos como factor y el CI como variable dependiente) se obtiene $F = 2,60$, con $p = 0,115$. Por tanto, no puede rechazarse la hipótesis nula de igualdad de medias y, consecuentemente, no hay razón para pensar que el factor (los métodos) esté afectando a la covariable (el CI).
3. *Las pendientes de regresión son homogéneas,* es decir, las pendientes que relacionan la covariable con la variable dependiente son iguales en todos los grupos definidos por los niveles del factor. Un análisis de regresión del rendimiento sobre el CI dentro de cada grupo arroja los siguientes coeficientes de regresión: 0,081, 0,079 y 0,113 (el coeficiente de regresión global, es decir, el que se obtiene con todos los casos, vale 0,108). Con estos resultados, ¿es razonable asumir que las tres pendientes de regresión poblacionales son iguales? Esto puede comprobarse ajustando un modelo que, además de los efectos individuales del factor y de la covariable, incluya el efecto de la interacción entre ambos.

Para ajustar este modelo con el SPSS: (1) reproducir los datos de la Tabla 2.1 en el *Editor de datos* o descargar el archivo *Motivación rendimiento* de la página web del manual; (2) seleccionar la opción **Modelo lineal general > Univariante** del menú **Analizar** para acceder al cuadro de diálogo *Univariante*; trasladar la variable *rendimiento* al cuadro **Dependiente**, la variable *método* a la lista **Factores fijos** y la variable *CI* (cociente intelectual) a la lista **Covariables**; (3) pulsar el botón **Modelo** para acceder al subcuadro de diálogo *Univariante: Modelo*, seleccionar la opción **Personalizado** y trasladar a la lista **Modelo** el término individual *método*, el término individual *CI* y la interacción *método*CI*; pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones se obtiene, entre otros resultados, una tabla resumen con una valoración de los efectos solicitados. De esta tabla únicamente nos interesa la información relativa a la interacción entre el factor (método) y la covariable (CI); el resto de los efectos los evaluaremos más adelante sin incluir el efecto de esta interacción. El efecto de la interacción tiene asociado un estadístico $F = 0,45$ con un nivel crítico ($\text{sig.} = 0,653$) mayor que 0,05. Este resultado indica que no existe evidencia de interacción entre la covariable y el factor; consecuentemente, no hay razón para pensar que las pendientes de regresión son distintas.

Resumiendo: del chequeo que acabamos de realizar se desprende que los datos de la Tabla 2.1 reúnen las condiciones necesarias para poder aplicar un modelo de ANCOVA: (1) la covariable está linealmente relacionada con la variable dependiente, (2) no hay evidencia de que el factor esté afectando a la covariable y (3) no hay evidencia de que la dirección o la intensidad de la relación entre la covariable y la variable dependiente cambie de un grupo a otro.

Cómo valorar el efecto del factor

El objetivo del análisis de covarianza que estamos llevando a cabo es valorar el efecto del factor (los métodos) sobre la variable dependiente (el rendimiento) controlando el efecto de la covariable (el CI). No obstante, comenzaremos valorando el efecto del factor prescindiendo de la covariable.

La finalidad de este análisis preliminar es poder contar con un punto de referencia con el que comparar los resultados del ANCOVA. Y consiste simplemente en aplicar un ANOVA con la variable *método* como factor y el *rendimiento* como variable dependiente. La Tabla 2.2 muestra los resultados de este análisis preliminar. Puesto que el nivel crítico obtenido ($\text{sig.} = 0,020$) es menor que 0,05, lo razonable es rechazar la hipótesis nula de igualdad de medias y concluir que el rendimiento medio no es el mismo con los tres métodos.

Tabla 2.2. Tabla resumen del ANOVA

Rendimiento					
	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	17,73	2	8,87	5,54	,020
Intra-grupos	19,20	12	1,60		
Total	36,93	14			

Al incorporar una covariable al análisis puede ocurrir que los resultados del ANOVA y los del ANCOVA sean iguales o puede ocurrir que sean distintos. Serán iguales cuando la presencia de la covariable no altere la relación entre el factor y la variable dependiente; serán distintos cuando la presencia de la covariable altere esa relación. En este segundo caso pueden estar ocurriendo dos cosas distintas: que un efecto significativo en ANOVA no lo sea en ANCOVA o que un efecto no significativo en ANOVA lo sea en ANCOVA. Lo primero (efecto significativo que deja de serlo) puede ocurrir porque la relación entre el factor y la variable dependiente es *espuria* y, eliminado el efecto de la covariable, al factor no le queda nada que explicar. Lo segundo (efecto no significativo que pasa a serlo) ocurre cuando el factor no está relacionado con la variable dependiente, pero sí con la parte de la variable dependiente que queda tras eliminar el efecto debido a la covariable.

Para valorar el efecto de los *métodos* controlando el efecto del *cociente intelectual*, es decir, para ajustar un modelo de ANCOVA a los datos de la Tabla 2.1:

- Reproducir los datos de la Tabla 2.1 en el *Editor de datos* o descargar el archivo *Motivación rendimiento* de la página web del manual.
- Seleccionar la opción **Modelo lineal general > Univariante** del menú **Analizar** para acceder al cuadro de diálogo *Univariante*.
- Trasladar la variable *rendimiento* al cuadro **Dependiente**, la variable *método* a la lista **Factores fijos** y la variable *CI* (cociente intelectual) a la lista **Covariables**.

Aceptando estas elecciones se obtienen, entre otros, los resultados⁸ que muestra la Tabla 2.3. El cociente intelectual (es decir, la covariable) está relacionada con el rendimiento ($F = 34,90$, $p < 0,05$; esto era lo esperable atendiendo a los resultados obtenidos al chequear los supuestos). Y el efecto de los métodos, que es el efecto que realmente interesa valorar, ha dejado de ser significativo ($F = 2,48$, $p = 0,129$).

Por tanto, cuando la covariable CI no interviene en el análisis (Tabla 2.2), el rendimiento medio parece no ser el mismo con los tres métodos; sin embargo, cuando interviene la covariable (Tabla 2.3), las diferencias en el rendimiento medio desaparecen. Por tanto, cuando se elimina del rendimiento la variabilidad atribuible al CI, los diferentes métodos no ayudan a entender o explicar las diferencias en el rendimiento.

Tabla 2.3. Tabla resumen del ANCOVA

Variable dependiente: Rendimiento

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.
Modelo corregido	32,33 ^a	3	10,78	25,76	,000
Intersección	2,03	1	2,03	4,86	,050
CI	14,60	1	14,60	34,90	,000
Método	2,08	2	1,04	2,48	,129
Error	4,60	11	,42		
Total	565,00	15			
Total corregida	36,93	14			

a. R cuadrado = ,875 (R cuadrado corregida = ,841)

Los datos de la Tabla 2.1 los hemos analizado mediante la comparación de dos modelos alternativos a los que hemos llamado modelo 0 y modelo 1 (ver ecuaciones [2.21] y [2.22]). La peculiaridad (y la utilidad) de estos dos modelos está en que únicamente difieren en el término que interesa valorar, es decir, en el término referido al efecto del factor. La información que ofrece la Tabla 2.3 se basa en la comparación de estos dos modelos.

Para realizar esta comparación se comienza estimando los parámetros μ , μ_j y β , y obteniendo los pronósticos y los residuos que se derivan de ambos modelos. Las mejores estimaciones de las medias poblacionales son las correspondientes medias muestrales:

$$\hat{\mu} = \bar{Y} = 5,93; \quad \hat{\mu}_1 = \bar{Y}_1 = 7,40; \quad \hat{\mu}_2 = \bar{Y}_2 = 5,60; \quad \hat{\mu}_3 = \bar{Y}_3 = 4,80$$

Y las mejores estimaciones del coeficiente de regresión β se consiguen mediante [2.30] para el modelo 0 y mediante [2.31] para el modelo 1. Con estas ecuaciones se obtiene

⁸ El procedimiento **Univariate** ya se ha explicado con detalle en el Capítulo 7 del segundo volumen. En este momento únicamente nos detendremos a explicar los aspectos relacionados con el nuevo elemento: la covariable. Para realizar comparaciones múltiples, estimar el tamaño del efecto, calcular la potencia observada, etc., sirve todo lo ya dicho en ese capítulo.

Tabla 2.4. Cálculos basados en los datos de la Tabla 2.1 (A = método; X = Cl; Y = rendimiento)

A	Y	X	$x_{ij(0)}$	$x_{ij(1)}$	$\hat{\mu}_{ij(0)}$	$\hat{E}_{ij(0)}$	$\hat{E}_{ij(0)}^2$	$\hat{\mu}_{ij(1)}$	$\hat{E}_{ij(1)}$	$\hat{E}_{ij(1)}^2$
1	6	100	-6	-16	5,29	0,71	0,51	5,97	0,03	0,00
1	8	130	24	14	8,53	-0,53	0,28	8,65	-0,65	0,43
1	7	110	4	-6	6,37	0,63	0,40	6,86	0,14	0,02
1	9	130	24	14	8,53	0,47	0,23	8,65	0,35	0,12
1	7	110	4	-6	6,37	0,63	0,40	6,86	0,14	0,02
2	4	90	-16	-13	4,21	-0,21	0,04	4,44	-0,44	0,19
2	7	120	14	17	7,45	-0,45	0,20	7,12	-0,12	0,02
2	6	95	-11	-8	4,75	1,25	1,57	4,88	1,12	1,25
2	5	100	-6	-3	5,29	-0,29	0,08	5,33	-0,33	0,11
2	6	110	4	7	6,37	-0,37	0,13	6,23	-0,23	0,05
3	3	80	-26	-19	3,13	-0,13	0,02	3,10	-0,10	0,01
3	5	100	-6	1	5,29	-0,29	0,08	4,89	0,11	0,01
3	7	110	4	11	6,37	0,63	0,40	5,79	1,21	1,48
3	4	100	-6	1	5,29	-1,29	1,65	4,89	-0,89	0,79
3	5	105	-1	6	5,83	-0,83	0,68	5,34	-0,34	0,11
<i>Suma</i>							6,68		4,60	

$\hat{\beta}_{(0)} = 0,10766$ para el modelo 0 y $\hat{\beta}_{(1)} = 0,08956$ para el modelo 1 (utilizamos 5 decimales para evitar los problemas derivados del redondeo).

Una vez que se tienen las estimaciones de las medias y de las pendientes, los pronósticos de cada modelo se obtienen aplicando [2.32] y [2.33]. Puesto que el modelo 0 no incluye el término relativo al efecto del factor (el modelo 0 no hace distinción entre grupos), basta una ecuación para obtener todos los pronósticos. Para obtener los pronósticos del modelo 1 hacen falta tres ecuaciones, una por grupo:

$$\text{Pronósticos del modelo 0: } \hat{\mu}_{ij(0)} = \hat{\mu} + \hat{\beta}_{(0)} x_{ij} = 5,93 + 0,10766(x_{ij})$$

$$\text{Pronósticos del modelo 1: } \hat{\mu}_{i1(1)} = \hat{\mu}_1 + \hat{\beta}_{(1)} x_{i1} = 7,40 + 0,08956(x_{i1})$$

$$\hat{\mu}_{i2(1)} = \hat{\mu}_2 + \hat{\beta}_{(1)} x_{i2} = 5,60 + 0,08956(x_{i2})$$

$$\hat{\mu}_{i3(1)} = \hat{\mu}_3 + \hat{\beta}_{(1)} x_{i3} = 4,80 + 0,08956(x_{i3})$$

Con estas ecuaciones se obtienen los pronósticos que recoge la Tabla 2.4 en las columnas $\hat{\mu}_{ij(0)}$ y $\hat{\mu}_{ij(1)}$ (los cálculos de esta tabla no es necesario, ni tampoco útil, hacerlos a mano; pueden obtenerse fácilmente con un programa informático como el SPSS).

Restando estos pronósticos a los valores observados (Y) se obtienen, tal como se indica en [2.34] y [2.35], los residuos de cada modelo. La Tabla 2.4 recoge estos re-

siduos en las columnas $\hat{E}_{ij(0)}$ y $\hat{E}_{ij(1)}$. Sumando ahora los cuadrados de los residuos se obtienen las correspondientes desviaciones (ver [2.36] y [2.37]):

$$-2LL_0 = \sum_i \sum_j \hat{E}_{ij(0)}^2 = 6,68$$

$$-2LL_1 = \sum_i \sum_j \hat{E}_{ij(1)}^2 = 4,60$$

Estas desviaciones indican cuánto se aleja cada modelo del ajuste perfecto, es decir, indican el grado de desajuste de cada modelo (cuanto mayor es la desviación, mayor es el desajuste). La diferencia entre ambas desviaciones indica cuánto se reduce el desajuste del modelo 0 al incorporar el término α_j , es decir, al incorporar el único término en el que difieren ambos modelos (esta diferencia es justamente la suma de cuadrados asociada al efecto del factor en la Tabla 2.3: $6,68 - 4,60 = 2,08$). Lo que hace el estadístico F definido en [2.38] es valorar esta diferencia entre las desviaciones.

Para obtener el estadístico F necesitamos, además de las desviaciones, los grados de libertad de cada modelo. En el modelo 0 se están estimando solo 2 parámetros (la media total y la pendiente de regresión); en el modelo 1 se están estimando 4 parámetros (las medias de los 3 grupos y la pendiente de regresión). Por tanto, en el modelo 0 se pierden 2 grados de libertad y en el modelo 1 se pierden 4. En consecuencia, el modelo 0 tiene $15 - 2 = 13$ grados de libertad y el modelo 1 tiene $15 - 4 = 11$ grados de libertad (15 es el número total de observaciones). Con las desviaciones que acabamos de calcular y con estos grados de libertad obtenemos

$$F = \frac{[-2LL_0 - (-2LL_1)] / (gl_0 - gl_1)}{-2LL_0 / gl_1} = \frac{(6,68 - 4,60) / 2}{4,60 / 11} = \frac{1,04}{0,42} = 2,48$$

que es justamente el valor que ofrece el SPSS (ver Tabla 2.3) para el estadístico F asociado al efecto del factor A (los métodos).

La Tabla 2.3 contiene toda la información necesaria para valorar el efecto del factor (los métodos) tras controlar el efecto de la covariable (el CI). Pero existe información adicional que puede ayudarnos a entender mejor lo que realmente se está haciendo con un modelo de ANCOVA.

Al ajustar un modelo de ANOVA nos estamos preguntando cuál es el efecto del factor. Al ajustar un modelo de ANCOVA nos estamos preguntando cuál sería el efecto del factor *si todos los grupos tuvieran la misma media en la covariable*. Esto significa que las medias que realmente se están comparando en un ANCOVA no son las medias *originales*, sino otras llamadas **medias corregidas** (ver [2.24]). Estas otras medias pueden estimarse mediante

$$\bar{Y}_j^* = \bar{Y}_j - \hat{\beta}_{(1)}(\bar{X}_j - \bar{X}) \quad [2.39]$$

En realidad se trata de *medias condicionales*: son las medias que se estima que corresponden a cada grupo en la variable dependiente cuando la covariable toma su valor me-

dio. En nuestro ejemplo, estas medias corregidas reflejan el rendimiento medio de cada grupo cuando el CI vale 106. Aplicando [2.39] a nuestros datos obtenemos:

$$\bar{Y}_1^* = 7,40 - 0,08956 (116 - 106) = 6,50$$

$$\bar{Y}_2^* = 5,60 - 0,08956 (113 - 106) = 5,87$$

$$\bar{Y}_3^* = 4,80 - 0,08956 (99 - 106) = 5,43$$

El SPSS ofrece estas medias corregidas al solicitar comparaciones entre los niveles del factor con la opción **Comparar efectos principales** del subcuadro de diálogo *Opciones*.

Pendientes de regresión heterogéneas

Si no puede asumirse que las pendientes de regresión son iguales, el modelo de ANCOVA debe reflejar esta circunstancia incorporando la posibilidad de estimar pendientes separadas para cada grupo. Para ello, el modelo [2.22] debe reformularse de la siguiente manera:

$$\text{Modelo 1: } Y_{ij} = \mu + \alpha_j + \beta_j x_{ij} + \epsilon_{ij(1)} \quad [2.40]$$

La única diferencia entre los modelos [2.22] y [2.40] está en la pendiente de regresión: en [2.22] se está trabajando con una sola pendiente (β); en [2.40] se está trabajando con tantas pendientes como grupos (β_j). En el segundo caso, el efecto de la covariable está *anidado* en el efecto del factor.

En lo que tiene que ver con la valoración del efecto del factor, ajustar el modelo propuesto en [2.40] equivale a ajustar el modelo que, además del efecto del factor y del efecto de la covariable, incluye el efecto de la interacción entre ambos. Y esto puede hacerse tal como ya hemos explicado anteriormente en el punto 3 del apartado *Cómo chequear los supuestos*.

Esta estrategia basada en el procedimiento **Univariante** también permite obtener (mediante la opción **Estimaciones de los parámetros** del subcuadro de diálogo *Opciones*) las estimaciones de las pendientes de regresión dentro de cada grupo, pero en el formato típico de SPSS, es decir, fijando en cero el último parámetro y estimando los demás por referencia a él.

Hay formas más rápidas de obtener estas estimaciones. En primer lugar, también con el procedimiento **Univariante**, pero ajustando un modelo personalizado que incluya el *efecto del factor* y el de la *interacción del factor con la covariable* (es decir, dejando fuera el *efecto de la covariable*; esta es la forma de indicar en el SPSS que el efecto de la covariable está anidado en el del factor; también se puede hacer esto mediante *sin-taxis*, pero no es necesario). En segundo lugar, mediante el procedimiento **Regresión lineal**, segmentando previamente el archivo de datos con la variable *factor* para poder obtener una ecuación para cada grupo. En el próximo capítulo estudiaremos con más detalle cómo ajustar modelos de regresión cuando se sospecha que las pendientes cambian dependiendo del grupo en el que se calculan.

Análisis de regresión lineal

El análisis de regresión lineal ya lo hemos abordado en el Capítulo 10 del segundo volumen prestando atención a los aspectos más relevantes del análisis. De hecho, ya hemos visto cómo llevar a cabo las cuatro tareas básicas relacionadas con el ajuste de un modelo lineal:

1. **Seleccionar el modelo.** Hemos visto cómo formular un modelo de regresión y cómo ajustarlo en un único paso o cómo proceder por pasos para obtener el máximo ajuste con el menor número de variables independientes. También hemos visto cómo incluir en el modelo variables independientes categóricas.
2. **Estimar los parámetros y obtener los pronósticos.** Hemos podido comprobar que las estimaciones de los coeficientes de un modelo de regresión lineal se basan en el criterio de mínimos cuadrados. También hemos visto cómo obtener pronósticos y cómo calcular intervalos de confianza para los pronósticos individuales y para los pronósticos promedio.
3. **Valorar la calidad o ajuste del modelo.** Para valorar el ajuste global hemos utilizado los estadísticos F (para la significación estadística) y R^2 (para la significación sustantiva). Para valorar la contribución de cada variable independiente hemos utilizado el estadístico T (para la significación estadística) y el cuadrado del coeficiente de correlación semiparcial (para la importancia relativa de cada variable).
4. **Chequear los supuestos.** Finalmente, hemos enumerado los supuestos de la regresión lineal y la forma de chequearlos. Y también hemos visto cómo detectar casos atípicos e influyentes.

Con lo ya estudiado en el Capítulo 10 del segundo volumen tenemos todo lo necesario para ajustar e interpretar con solvencia un modelo de regresión lineal. Por tanto, no repetiremos aquí lo que ya está dicho allí. Nos limitaremos a repasar brevemente el análisis de regresión lineal desde la perspectiva de la modelización lineal y prestaremos atención a un aspecto todavía no tratado: la interacción entre variables independientes.

Seleccionar el modelo

Al igual que ocurre al ajustar un modelo de análisis de varianza o de covarianza, al ajustar un modelo de regresión lineal se están planteando dos modelos rivales o alternativos: el **modelo nulo** (modelo 0), que, aparte de los errores aleatorios, únicamente incluye el término constante o intersección y el **modelo propuesto** (modelo 1), que, además de los errores aleatorios, incluye todos los efectos tenidos en cuenta:

$$\text{Modelo 0: } Y_i = \beta_{0(0)} + E_{i(0)} \quad [2.41]$$

$$\text{Modelo 1: } Y_i = \beta_{0(1)} + \beta_1 X_i + E_{i(1)} \quad [2.42]$$

(el subíndice i se refiere a los sujetos: $i = 1, 2, \dots, n$; los subíndices 0 y 1 entre paréntesis indican de qué modelo se trata). El significado de los términos que incluyen estos modelos ya se ha explicado en el capítulo anterior a propósito de la ecuación [1.1].

La ecuación [2.42] es un modelo de regresión simple (una sola variable independiente). Añadiendo variables independientes a la ecuación se puede construir un modelo de regresión múltiple y trabajar con la misma lógica que utilizaremos con ésta.

Puesto que los errores no intervienen en los pronósticos, con el modelo 0 se está asignando el mismo pronóstico a todos los sujetos, mientras que con el modelo 1 se está asignando un pronóstico distinto a cada patrón de variabilidad:

$$\mu_{i(0)} = \beta_{0(0)} \quad [2.43]$$

$$\mu_{i(1)} = \beta_{0(1)} + \beta_1 X_i \quad [2.44]$$

Los errores aleatorios son, en ambos casos, las diferencias entre los valores observados y los pronosticados. Por tanto,

$$E_{i(0)} = Y_i - \mu_{i(0)} \quad [2.45]$$

$$E_{i(1)} = Y_i - \mu_{i(1)} \quad [2.46]$$

Cuando no existe una idea clara acerca de qué modelo concreto ajustar, es decir, cuando no se tiene una hipótesis concreta acerca de qué variables independientes pueden ayudar a explicar o entender el comportamiento de la variable dependiente, en lugar de proponer un modelo concreto, puede procederse por pasos para encontrar el modelo capaz de ofrecer el mejor ajuste posible con el menor número de variables. En el Capítulo 10 del segundo volumen se explica la regresión lineal jerárquica o por pasos.

Estimar los parámetros y obtener los pronósticos

El objetivo principal del análisis es averiguar si el término extra que incluye el modelo 1 permite mejorar el ajuste del modelo 0. Alcanzar este objetivo pasa por comparar el ajuste de ambos modelos. Ahora bien, la calidad de un modelo lineal se evalúa comparando los valores observados y los pronosticados. Y para obtener los pronósticos es necesario estimar los parámetros.

Aunque estas estimaciones pueden obtenerse aplicando diferentes métodos, en los modelos de regresión lineal suele utilizarse el método de **mínimos cuadrados**. Este método utiliza como estimadores de los parámetros los valores que minimizan la suma de las diferencias al cuadrado entre los valores observados y los pronosticados. En el caso de los parámetros incluidos en los modelos [2.43] y [2.44], los estimadores que ofrece el método de mínimos cuadrados son los siguientes:

$$\begin{aligned} \hat{\beta}_{0(0)} &= \bar{Y} \\ \hat{\beta}_{0(1)} &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 &= S_{XY} / S_X^2 \end{aligned} \quad [2.47]$$

Los pronósticos que se derivan de los modelos 0 y 1 se obtienen simplemente sustituyendo los parámetros de [2.43] y [2.44] por sus correspondientes estimadores:

$$\text{Pronósticos del modelo 0: } \hat{\mu}_{i(0)} = \bar{Y} \quad [2.48]$$

$$\text{Pronósticos del modelo 1: } \hat{\mu}_{i(1)} = \bar{Y} - \hat{\beta}_1 \bar{X} + (S_{XY}/S_X^2) X_i \quad [2.49]$$

La Tabla 2.5 recoge los datos de una muestra de 20 pacientes con trastorno depresivo que han participado en un estudio diseñado para valorar la eficacia de dos tratamientos distintos. Son los mismos datos ya analizados en el Capítulo 10 del segundo volumen. El estudio comenzó administrando la Escala de Depresión de Hamilton para obtener una medida inicial (*basal*) del nivel de depresión de los pacientes. Al finalizar el tratamiento se volvió a administrar la escala y se obtuvo una medida de la recuperación (*recuperac.*) restando las puntuaciones finales a las basales (los datos se encuentran en el archivo *Depresión hamilton reducido*, en la página web del manual).

Tabla 2.5. Datos de 20 pacientes sometidos a tratamiento antidepresivo

<i>id</i>	<i>basal</i>	<i>recuperac.</i>	$\hat{\mu}_{i(0)}$	$\hat{E}_{i(0)}$	$\hat{E}_{i(0)}^2$	$\hat{\mu}_{i(1)}$	$\hat{E}_{i(1)}$	$\hat{E}_{i(1)}^2$
1	25	5	9,95	-4,95	24,50	7,49	-2,49	6,21
2	23	5	9,95	-4,95	24,50	6,32	-1,32	1,75
3	21	2	9,95	-7,95	63,20	5,15	-3,15	9,93
4	22	8	9,95	-1,95	3,80	5,74	2,26	5,12
5	35	8	9,95	-1,95	3,80	13,34	-5,34	28,56
6	28	6	9,95	-3,95	15,60	9,25	-3,25	10,55
7	36	11	9,95	1,05	1,10	13,93	-2,93	8,58
8	30	6	9,95	-3,95	15,60	10,42	-4,42	19,52
9	27	9	9,95	0,95	0,90	8,66	0,34	0,11
10	29	8	9,95	-1,95	3,80	9,83	-1,83	3,36
11	32	12	9,95	2,05	4,20	11,59	0,41	0,17
12	27	12	9,95	2,05	4,20	8,66	3,34	11,14
13	30	11	9,95	1,05	1,10	10,42	0,58	0,34
14	32	16	9,95	6,05	36,60	11,59	4,41	19,46
15	27	10	9,95	0,05	0,00	8,66	1,34	1,79
16	25	9	9,95	-0,95	0,90	7,49	1,51	2,27
17	35	13	9,95	3,05	9,30	13,34	-0,34	0,12
18	38	16	9,95	6,05	36,60	15,10	0,90	0,81
19	34	18	9,95	8,05	64,80	12,76	5,24	27,47
20	28	14	9,95	4,05	16,40	9,25	4,75	22,58
	584	199			330,95			179,85

La primera columna de la tabla muestra el número de caso. Las dos siguientes columnas contienen los datos (la tabla de datos propuesta en el Capítulo 10 del segundo volumen incluye más variables que la tabla de datos que estamos proponiendo ahora; aquí únicamente hemos incluido las dos variables que vamos a utilizar en nuestro ejemplo: *basal* y *recuperación*). Las seis restantes columnas de la tabla recogen los cálculos realizados para ajustar los modelos [2.43] y [2.44].

Tomando la *recuperación* como variable dependiente y las puntuaciones basales (*basal*) como variable independiente, las ecuaciones propuestas en [2.47] ofrecen las siguientes estimaciones:

$$\hat{\beta}_{0(0)} = \bar{Y} = 9,95$$

$$\hat{\beta}_{0(1)} = \bar{Y} - \hat{\beta}_1 \bar{X} = 9,95 - 0,585(29,20) = -7,13$$

$$\hat{\beta}_1 = s_{XY}/s_X^2 = 13,59/23,22 = 0,585$$

Sustituyendo ahora los parámetros de [2.43] y [2.44] por sus correspondientes estimaciones se obtienen los pronósticos que recogen las columnas $\hat{\mu}_{i(0)}$ y $\hat{\mu}_{i(1)}$ de la Tabla 2.5. Utilizar esta aproximación basada en la comparación de modelos permite constatar que el modelo 0 pronostica, efectivamente, el mismo valor a todos los casos (la media de Y), mientras que el modelo 1 pronostica un valor distinto para cada patrón de variabilidad (en el ejemplo, un valor distinto para cada puntuación basal distinta).

Valorar la calidad o ajuste del modelo

La comparación entre modelos rivales se basa en la **desvianza** ($-2LL$), la cual ya sabemos que se obtiene a partir de los residuos, es decir, a partir de las diferencias entre los valores *observados* (Y_i) y los *pronosticados* ($\hat{\mu}_i$):

$$\text{Residuos del modelo 0: } \hat{E}_{i(0)} = Y_i - \hat{\mu}_{i(0)} \quad [2.50]$$

$$\text{Residuos del modelo 1: } \hat{E}_{i(1)} = Y_i - \hat{\mu}_{i(1)} \quad [2.51]$$

El estadístico $-2LL$ se obtiene a partir de la suma de los cuadrados de estos residuos (la Tabla 2.5 recoge estas sumas de cuadrados en la última fila):

$$-2LL_0 = \sum_i \hat{E}_{i(0)}^2 = 330,95 \quad [2.52]$$

$$-2LL_1 = \sum_i \hat{E}_{i(1)}^2 = 179,85 \quad [2.53]$$

La desvianza refleja el grado de *desajuste* de un modelo, es decir, el grado en que un modelo se aleja del ajuste perfecto. La desvianza del modelo 0 ($-2LL_0$), que en los modelos de regresión lineal es la *suma de cuadrados total*, refleja el máximo desajuste posible (el desajuste que se deriva de pronosticar la variable dependiente sin otra información que la propia variable dependiente). La desvianza del modelo 1 ($-2LL_1$), que en los modelos de regresión es la *suma de cuadrados error* o *residual*, refleja el desa-

juste del modelo que incorpora la información de la variable independiente. Por tanto, la diferencia entre ambas desviaciones refleja la diferencia en el desajuste de ambos modelos. Dividiendo esa diferencia entre la desviación del modelo 0 se obtiene el estadístico R^2 (el coeficiente de determinación), el cual expresa la proporción en que el modelo 1 reduce el desajuste del modelo 0:

$$R^2 = \frac{-2LL_0 - (-2LL_1)}{-2LL_0} = \frac{330,95 - 179,85}{330,95} = 0,46 \quad [2.54]$$

Este resultado indica que las puntuaciones basales (única variable independiente que incluye el modelo de regresión que estamos ajustando) consigue reducir el desajuste del modelo nulo un 46%.

Si la diferencia entre las desviaciones de los modelos 0 y 1 se divide, no entre la desviación del modelo 0, sino entre la desviación del modelo 1, se obtiene una estimación de la *proporción en que aumenta el desajuste (PAD)* al eliminar del modelo 1 la variable independiente:

$$PAD = \frac{-2LL_0 - (-2LL_1)}{-2LL_1} = \frac{330,95 - 179,85}{179,85} = 0,84 \quad [2.55]$$

Este resultado indica que, al eliminar las puntuaciones basales de nuestro modelo de regresión, el desajuste aumenta un 84%.

Cuanto mayor es el valor del estadístico *PAD*, mayor es la diferencia en el desajuste de los modelos 0 y 1. Dividiendo el numerador y el denominador de [2.53] entre sus respectivos grados de libertad (gl) se obtiene un estadístico que permite valorar la diferencia entre el desajuste de ambos modelos:

$$F = \frac{[-2LL_0 - (-2LL_1)] / (gl_0 - gl_1)}{-2LL_1 / gl_1} \quad [2.56]$$

Bajo ciertas condiciones (ver siguiente apartado), la distribución muestral de este estadístico se aproxima a la distribución F con p y $n - p - 1$ grados de libertad⁹ (p se refiere al número de variables independientes y n al número de casos). En nuestro ejemplo, con $n = 20$ y $p = 1$,

$$F = \frac{(330,95 - 179,85) / 1}{179,85 / 18} = 15,12$$

El estadístico F permite contrastar la hipótesis nula de que el término extra que incluye el modelo 1 vale cero en la población (es decir, la hipótesis nula de que β_1 vale cero).

⁹ Los grados de libertad de un modelo de regresión lineal se obtienen restando al número de casos (n) el número de parámetros estimados. Los grados de libertad del modelo 0 son $n - 1$: puesto que solo se estima el término constante, solo se pierde un grado de libertad. Los grados de libertad del modelo 1 son $n - p - 1$: se pierde un grado de libertad por el término constante y uno más por cada variable independiente.

Un estadístico F significativo ($p < 0,05$) permitirá rechazar esta hipótesis nula y concluir que la variable independiente está relacionada con la variable dependiente. La probabilidad de encontrar valores mayores que 15,12 en la distribución F con 1 y 18 grados de libertad vale 0,001 (este resultado puede obtenerse en SPSS con la función CDF.F de la opción **Calcular**). Por tanto, podemos concluir que al eliminar del modelo 1 las puntuaciones basales, se produce un aumento significativo del desajuste.

Chequear los supuestos

Para que un modelo lineal funcione correctamente es necesario que se den ciertas condiciones. Estas condiciones son las que garantizan que las estimaciones que realizamos son insesgadas y eficientes, y que el estadístico F funciona como se espera que lo haga. En el caso del modelo de regresión lineal, estas condiciones son las que ya hemos llamado abreviadamente *linealidad*, *independencia*, *normalidad*, *igualdad de varianzas* y *no colinealidad*. En el Capítulo 10 del segundo volumen se explica todo lo relativo a estos supuestos y a la forma de chequearlos.

Interacción entre variables independientes

En las ecuaciones de regresión estudiadas hasta ahora nos hemos limitado a pronosticar la variable dependiente Y a partir de una única variable independiente o a partir de varias variables independientes combinadas de forma aditiva (sumadas).

Combinar las variables independientes de forma aditiva implica asumir que no interaccionan entre sí. En una ecuación de estas características, *el pronóstico para Y cambia de forma constante con cada unidad que aumenta cada variable independiente, cualquiera que sea el valor que tomen el resto de variables independientes presentes en la ecuación*.

Cuando la relación entre una variable dependiente (por ejemplo, la *recuperación*) y una variable independiente (por ejemplo, el *tto*) cambia en función de los valores que toma una tercera variable (por ejemplo, la *edad*), combinar aditivamente en una ecuación el *tto* y la *edad* no permite entender lo que está ocurriendo. Si dos variables independientes interaccionan, la ecuación de regresión debe incluir un término adicional para reflejar esa circunstancia¹⁰.

Para incorporar a un modelo de regresión el efecto debido a la interacción entre variables independientes basta con incluir el *producto de las variables que interaccionan*. Un modelo de regresión lineal **aditivo**, con dos variables independientes (X_1 y X_2), adopta la siguiente forma:

$$\mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \quad [2.57]$$

¹⁰ Para profundizar en todo lo relativo a la interpretación de las interacciones en un modelo de regresión lineal puede consultarse Jaccard y Turrisi (2003).

El correspondiente modelo de regresión lineal **no aditivo** incluye, además de los términos del modelo aditivo [2.57], el producto entre X_1 y X_2 :

$$\mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} \quad [2.58]$$

En este capítulo y en el 10 del segundo volumen hemos estudiado ya lo relativo al ajuste de modelos aditivos como el propuesto en [2.57]. Para estimar con el SPSS un modelo no aditivo como el propuesto en [2.58] basta con trasladar, en el cuadro de diálogo principal (opción **Regresión > Lineal** del menú **Analizar**), la variable dependiente (Y) al cuadro **Dependiente** y las variables independientes (X_1 y X_2) y el producto entre ambas ($X_1 X_2$) a la lista **Independientes**. Para poder utilizar el producto $X_1 X_2$ en este procedimiento es necesario haberlo creado previamente (esto puede hacerse fácilmente con la opción **Calcular** del menú **Transformar**).

Al incorporar al modelo de regresión un término con la interacción $X_1 X_2$, no solo la situación se vuelve más compleja, sino que el significado de los coeficientes asociados a los efectos individuales cambia sensiblemente. Para facilitar la interpretación de los resultados vamos a considerar dos escenarios: (1) dos variables independientes cuantitativas y (2) una variable independiente categórica y la otra cuantitativa.

Dos variables cuantitativas

Retomemos nuestro ejemplo sobre la recuperación de pacientes sometidos a tratamiento antidepresivo (los datos se encuentran en el archivo *Depresión hamilton interacción*, el cual puede descargarse de la página web del manual). Una ecuación de regresión *no aditiva* con la *recuperación* como variable dependiente (con valor esperado μ_i) y las *puntuaciones basales* (X_1) y la *edad* (X_2) como variables independientes adopta la forma:

$$\mu_i = \beta_0 + \beta_1 (cbasal) + \beta_2 (cedad) + \beta_3 (cbasal \times cedad) \quad [2.59]$$

Recordemos que el coeficiente β_0 es la recuperación estimada cuando todas las variables independientes valen cero. Por tanto, β_0 solamente tiene significado si también lo tiene el valor cero de todas las variables independientes. Por este motivo, y también para facilitar después la interpretación del resto de coeficientes, en lugar de las variables originales *basal* y *edad* estamos utilizando las variables *cbasal* (puntuaciones basales *centradas*) y *cedad* (edad *centrada*). Ambas variables se han *centrado* tomando como referencia un valor próximo al centro de sus respectivas distribuciones: 30 en el caso de las puntuaciones basales y 50 en el caso de la edad. Por tanto, el valor *cbasal* = 0 se refiere a una puntuación basal de 30 puntos y el valor *cedad* = 0 se refiere a una edad de 50 años.

La Tabla 2.6 muestra los resultados obtenidos al estimar el modelo propuesto en [2.59]. La ecuación de regresión que ofrece la tabla es:

$$\hat{\mu}_i = 10,59 + 2,22 (cbasal) - 0,28 (cedad) - 0,03 (cbasal \times cedad)$$

Tabla 2.6. Variables incluidas en la ecuación (con la interacción *basal centrada* por *edad centrada*)

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	10,59	,55		19,42	,000
	<i>cbasal</i>	2,22	,85	2,57	2,61	,019
	<i>cedad</i>	-,28	,08	-,49	-3,69	,002
	<i>cbasal x cedad</i>	-,03	,02	-2,01	-2,03	,059

Aunque el efecto de la interacción *cbasal* × *cedad* no alcanza la significación estadística (*sig.* = 0,059), primero vamos a interpretar los coeficientes del modelo como si esta interacción fuera significativa y, a continuación, teniendo en cuenta que no lo es:

- **Coeficiente $\hat{\beta}_0$.** La constante es el pronóstico que ofrece la ecuación cuando todas las variables independientes valen cero. Puesto que las dos variables independientes que incluye la ecuación, *cbasal* y *cedad*, están centradas en 30 y 50, respectivamente, el valor de la constante (10,59) es la recuperación pronosticada para los pacientes que tienen una puntuación basal de 30 puntos y una edad de 50 años.
- **Coeficiente $\hat{\beta}_1$ (*cbasal*).** El coeficiente asociado a la variable *cbasal* recoge el efecto de esa variable cuando *cedad* = 0, es decir, el efecto de las puntuaciones basales entre los pacientes de 50 años. El valor del coeficiente (2,22) indica que, entre los pacientes que tienen 50 años, la recuperación pronosticada aumenta 2,22 puntos por cada punto que aumentan las puntuaciones basales.

Cuando una ecuación de regresión incluye una interacción entre variables independientes, el significado de los coeficientes de regresión asociados a los efectos individuales cambia de forma importante. En una ecuación que no incluye el producto X_1X_2 , el coeficiente $\hat{\beta}_1$ indica cómo cambia Y por cada unidad que aumenta X_1 , cualquiera que sea el valor de X_2 . En una ecuación que incluye el producto X_1X_2 , el coeficiente $\hat{\beta}_1$ sigue indicando cómo cambia Y por cada unidad que aumenta X_1 , pero no para cualquier valor de X_2 , sino solo para el valor $X_2 = 0$. Así es como hemos interpretado el coeficiente $\hat{\beta}_1$ en el párrafo anterior: la relación entre *cbasal* y *recuperación* la hemos referido a los pacientes que tienen 50 años (*cedad* = 0).

Ahora bien, si el efecto de la interacción no alcanza la significación estadística, los coeficientes asociados a los efectos individuales deben interpretarse como si el efecto de la interacción no hubiera sido incluido en el modelo. De hecho, los coeficientes de regresión no significativos corresponden a efectos que pueden eliminarse del modelo sin que el ajuste se vea alterado. En nuestro ejemplo, puesto que el efecto de la interacción no alcanza la significación estadística, el coeficiente asociado a la variable *cbasal* indica que la recuperación pronosticada aumenta 2,22 puntos por cada punto que aumentan las puntuaciones basales, *cualquiera que sea la edad de los pacientes*.

- **Coeficiente $\hat{\beta}_2$ (*cedad*).** El coeficiente asociado a la variable *cedad* recoge el efecto de esa variable cuando *cbasal* = 0, es decir, el efecto de *cedad* cuando la puntuación

basal vale 30 puntos. El valor del coeficiente ($-0,28$) indica que, entre los pacientes con una puntuación basal de 30, la recuperación pronosticada disminuye 0,28 puntos por cada año que aumenta la edad.

Esta sería la interpretación de $\hat{\beta}_2$ en presencia de una interacción significativa. Pero como el efecto de la interacción no alcanza la significación estadística, el coeficiente $\hat{\beta}_2$ debe interpretarse como si el efecto de la interacción no hubiera sido incluido en el modelo: la recuperación pronosticada disminuye 0,28 puntos por cada año que aumenta la edad, *cualquiera que sea la puntuación basal de los pacientes*.

- **Coeficiente $\hat{\beta}_3$ ($cbasal \times cedad$).** Por último, el coeficiente de regresión asociado al efecto de la interacción $cbasal \times cedad$ indica cómo va cambiando la relación entre la recuperación y las puntuaciones basales al ir aumentando la edad. El valor obtenido ($-0,03$) permite concretar que la pendiente que relaciona la recuperación con las puntuaciones basales va disminuyendo 0,03 puntos con cada año que va aumentando la edad. No obstante, este cambio de 0,03 puntos no alcanza la significación estadística ($sig. = 0,059$).

En el párrafo anterior se ha considerado que *cedad* actúa como variable *moderadora* de la relación entre *cbasal* y *recuperación*, pero el coeficiente $\hat{\beta}_3$ también puede interpretarse intercambiando el rol de las variables, es decir, tomando *cbasal* como variable moderadora de la relación entre *cedad* y *recuperación*: la pendiente que relaciona la recuperación con la edad va disminuyendo 0,03 puntos con cada unidad que aumentan las puntuaciones basales. Elegir una u otra interpretación es algo que depende, básicamente, de la justificación teórica que se tenga acerca de qué variable de las dos independientes es moderadora del efecto de la otra.

Una interacción no significativa puede ser eliminada del modelo sin que se resienta la calidad del mismo. Una interacción no significativa indica que no existe evidencia de que *cbasal* modere la relación entre *cedad* y *recuperación*, ni de que *cedad* modere la relación entre *cbasal* y *recuperación*.

Una variable dicotómica y una cuantitativa

Consideremos ahora una ecuación de regresión *no aditiva* con la variable *recuperación* (Y) como variable dependiente y las variables *tto* (X_1) y *edad* (X_2) como variables independientes:

$$\mu_i = \beta_0 + \beta_1(tto) + \beta_2(cedad) + \beta_3(tto \times cedad) \quad [2.60]$$

Para facilitar la interpretación de los coeficientes, en lugar de la variable original *edad* estamos utilizando la variable *cedad* (*edad centrada*), la cual se ha centrado restando 50 a todas las edades (por tanto, el valor $cedad = 0$ se refiere a los pacientes que tienen 50 años). En la variable *tto*, el código 0 corresponde al tratamiento *estándar* y el código 1 al *combinado*. La Tabla 2.7 muestra los resultados del análisis. La ecuación de regresión obtenida es:

$$\hat{\mu}_i = 6,49 + 6,72(tto) + 0,04(cedad) - 0,02(tto \times cedad)$$

Tabla 2.7. Variables incluidas en la ecuación (con la interacción *tto* por *edad centrada*)

Modelo	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
	B	Error típ.	Beta		
1 (Constante)	6,49	1,57		4,13	,001
tto	6,72	2,06	,83	3,27	,005
cedad	,04	,19	,08	,24	,815
tto x cedad	-,02	,30	-,01	-,06	,954

Aunque el efecto de la interacción *tto* × *cedad* es no significativo (*sig.* = 0,954), vamos a interpretar los coeficientes del modelo, primero, como si esta interacción fuera significativa y, después, teniendo en cuenta que no lo es:

- *Coeficiente $\hat{\beta}_0$* . La constante sigue siendo la recuperación que pronostica la ecuación de regresión cuando todas las variables independientes valen cero. Por tanto, el valor 6,49 es la recuperación pronosticada a los pacientes que tienen 50 años (*cedad* = 0) y que han recibido el tratamiento estándar (*tto* = 0).
- *Coeficiente $\hat{\beta}_1$ (*tto*)*. El coeficiente asociado a la variable *tto* recoge el efecto de esa variable cuando *cedad* = 0 (es decir, 50 años). El valor del coeficiente (6,72) indica que, entre los pacientes de 50 años, a los que han recibido el tratamiento combinado (*tto* = 1) se les pronostica una recuperación 6,72 puntos mayor que a los que han recibido el tratamiento estándar (*tto* = 0). En realidad, puesto que el efecto de la interacción *tto* × *cedad* es no significativo, la ventaja del tratamiento combinado sobre el estándar se refiere, no solo a los pacientes que tienen 50 años, sino a los pacientes de cualquier edad.
- *Coeficiente $\hat{\beta}_2$ (*cedad*)*. El coeficiente asociado a la variable *cedad* recoge el efecto de esa variable entre quienes han recibido el tratamiento estándar (*tto* = 0). El valor del coeficiente (0,04) indica que, entre los pacientes que han recibido el tratamiento estándar, la recuperación pronosticada va aumentando 0,06 puntos con cada año más de edad. No obstante, ese aumento de 0,06 puntos no es estadísticamente significativo (*sig.* = 0,890). Y, puesto que la interacción *tto* × *cedad* es no significativa, la falta de evidencia de relación entre la recuperación y la edad vale para ambos tratamientos.
- *Coeficiente $\hat{\beta}_3$ (*tto* × *cedad*)*. El coeficiente de regresión asociado al efecto de la interacción (-0,02) indica cómo va cambiando la relación estimada entre los *tratamientos* y la *recuperación* al ir aumentando la *edad*. En concreto, la pendiente que relaciona la recuperación con los tratamientos se estima que disminuye 0,02 puntos con cada año más de edad. Pero esta disminución es estadísticamente no significativa (*sig.* = 0,954).

En esta interpretación del efecto de la interacción hemos puesto el énfasis en la relación entre la variable independiente *tto* y la variable dependiente *recuperación*; es decir, hemos considerado que es la variable cuantitativa (*edad*) la que *mo-*

dera la relación entre las otras dos variables (los *tratamientos* y la *recuperación*). Esto es lo que, en principio, parece tener más sentido y por esta razón lo hemos hecho así. Pero, en el caso de que lo que tuviera sentido fuera lo contrario, estos mismos resultados pueden interpretarse asumiendo que la variable moderadora es la variable categórica (*tto*) y, por tanto, poniendo el énfasis de la interpretación en la relación entre la *edad* y la *recuperación*. En ese caso, lo que habría que concluir es que la pendiente que relaciona la *recuperación* con la *edad* es 0,02 puntos menor con el tratamiento *estándar* ($tto = 0$) que con el tratamiento *combinado* ($tto = 1$). Pero no debemos olvidar que esta diferencia es estadísticamente no significativa ($sig. = 0,954$).

Por supuesto, si una interacción es no significativa, lo razonable es asumir que su efecto es nulo y, consecuentemente con ello, no interpretarla; si la hemos interpretado aquí ha sido únicamente para explicar cómo se hace. Por otro lado, puesto que una interacción no significativa únicamente contribuye a complicar un modelo sin mejorar su ajuste, lo que debe hacerse con ella es simplemente eliminarla. En nuestro ejemplo, al eliminar la interacción $tto \times edad$, el coeficiente de determinación no se altera (vale 0,60 tanto si se incluye la interacción $tto \times edad$ como si no se incluye) y el coeficiente de determinación corregido no solo no disminuye sino que aumenta de 0,53 a 0,55.

Apéndice 2

Elementos de un modelo lineal clásico

¿Por qué para describir los datos correspondientes a un diseño de un factor utilizamos un modelo de las características del propuesto en [2.2]?

Supongamos que tenemos 3 muestras aleatorias de tamaño $n = 5$, cada una de las cuales ha recibido un tratamiento distinto ($J = 3$). Supongamos además que en cada sujeto hemos tomado una medida (Y_{ij}) relacionada con el efecto del tratamiento. Supongamos, por último, que se han obtenido los datos que muestra la Tabla 2.8.

Tabla 2.8. Ausencia de variabilidad ($\bar{Y} = 5$)

<i>Factor</i>	<i>Observaciones</i>					$\bar{Y}_j$
a_1	5	5	5	5	5	5
a_2	5	5	5	5	5	5
a_3	5	5	5	5	5	5

La peculiaridad de esta tabla es que las puntuaciones son iguales. No existe variabilidad ni entre los sujetos del mismo grupo ni entre las medias de los diferentes grupos. En este escenario, para describir correctamente lo que está ocurriendo basta con realizar un único pronóstico. Por tanto, los datos pueden describirse apropiadamente mediante un modelo que incluya un único parámetro (la media total μ):

$$Y_{ij} = \mu \quad [2.61]$$

Imaginemos ahora que, en lugar de los datos de la Tabla 2.8, obtenemos los datos que recoge la Tabla 2.9.

Tabla 2.9. Variabilidad entre los niveles del factor ($\bar{Y} = 5$)

<i>Factor</i>	<i>Observaciones</i>					$\bar{Y}_j$
a_1	2	2	2	2	2	2
a_2	6	6	6	6	6	6
a_3	7	7	7	7	7	7

Ahora, las medias de los grupos son distintas (variabilidad entre los grupos o *intergrupos*) pero todos los sujetos del mismo grupo siguen teniendo la misma puntuación. Para poder realizar pronósticos correctos en este nuevo escenario es necesario utilizar un modelo que, además de la media total (que todos los sujetos comparten), incorpore lo que cada grupo tiene de específico:

$$Y_{ij} = \mu + \alpha_j \quad [2.62]$$

Este modelo recoge, por un lado, la parte de Y que todos los sujetos tienen en común (μ) y, por otro, la parte de Y específica de cada grupo (α_j). Los datos de la Tabla 2.9 indican que lo que cada grupo tiene de específico es justamente su desviación de la media total; de ahí que el efecto asociado a cada tratamiento (α_j) se conciba e interprete como la diferencia entre la media de ese tratamiento y la media total: $\alpha_j = \mu_j - \mu$.

Pero ocurre que la realidad suele ser más compleja de lo que sugieren los datos de la Tabla 2.9. En el mundo real, además de variabilidad *entre* los grupos (*intergrupos*) también suele darse variabilidad *dentro* de los grupos (*intragrupos*). La Tabla 2.10 ofrece unos datos más parecidos a los que podrían obtenerse en un estudio real.

En este nuevo escenario, para poder pronosticar correctamente cada puntuación Y es necesario utilizar, además de μ y α_j , un nuevo término que refleje la variabilidad existente dentro de cada grupo:

$$Y_{ij} = \mu + \alpha_j + E_{ij} \quad [2.63]$$

Tabla 2.10 Variabilidad entre los niveles del factor y dentro de cada nivel ($\bar{Y} = 5$)

<i>Factor</i>	<i>Observaciones</i>					$\bar{Y}_j$
a_1	3	0	2	1	4	2
a_2	8	5	4	6	7	6
a_3	5	6	8	7	9	7

Así pues, para describir las puntuaciones Y correspondientes a J grupos aleatoriamente asignados a los J niveles de una variable independiente o factor, el modelo propuesto debe incluir tres términos: uno referido a la parte de Y que es común a todos los sujetos (la media total, μ), otro referido a la parte de Y que es específica de cada grupo (el efecto del factor, α_j) y otro más referido a la parte de Y que es específica de cada sujeto (los errores, E_{ij}).

Modelos lineales mixtos

Efectos fijos, aleatorios y mixtos

Los niveles o categorías de una variable independiente o factor pueden establecerse de dos maneras distintas:

1. Fijando los niveles que se desea estudiar (por ejemplo, cantidad de fármaco: 0 mg, 250 mg, 500 mg) o utilizando los niveles que posee el factor (por ejemplo, nivel educativo: sin estudios, primarios, secundarios, medios, superiores).
2. Seleccionando aleatoriamente unos pocos niveles de la población de posibles niveles del factor (por ejemplo, seleccionando una muestra aleatoria de los hospitales de una ciudad).

En el primer caso se tiene un factor de *efectos fijos*; en el segundo, un factor de *efectos aleatorios*. Los factores utilizados con mayor frecuencia en los modelos lineales son de efectos fijos. De hecho, en los capítulos sobre ANOVA incluidos en el segundo volumen (ver Capítulos 6 al 9) se ha puesto todo el énfasis en el estudio de factores de efectos fijos.

Sin embargo, no son infrecuentes las situaciones donde lo apropiado es utilizar factores de efectos aleatorios. Por ejemplo, para estudiar el tiempo de convalecencia tras una determinada intervención quirúrgica habrá que utilizar factores de efectos fijos como la gravedad de la enfermedad, el tipo de intervención, etc. Pero, probablemente, los pacientes habrá que seleccionarlos de distintos hospitales y este hecho no podrá pasarse por alto (pues la eficacia, la organización, etc., de todos los hospitales no es la

misma). Para estudiar el efecto del factor *hospital* podría seleccionarse aleatoriamente una muestra de hospitales (no sería necesario, ni tal vez posible, seleccionar todos los hospitales). Al proceder de esta manera, los resultados del estudio estarían indicando, no si dos hospitales concretos difieren entre sí (aquí no interesa averiguar si tal hospital concreto difiere de tal otro), sino si el factor *hospital* está relacionado con el tiempo de convalecencia posquirúrgica.

Un modelo lineal puede incluir, además de los términos correspondientes a los factores individualmente considerados, términos formados por la combinación de más de un factor, es decir, interacciones. Los términos (ya sean factores individuales o interacciones entre factores) que únicamente incluyen factores de efectos fijos se consideran términos de efectos fijos; los términos que incluyen factores de efectos aleatorios o una combinación de factores de efectos fijos y aleatorios se consideran términos de efectos aleatorios.

Qué es un modelo lineal mixto

A un modelo lineal que únicamente incluye términos de efectos fijos (al margen de los errores) se le llama *modelo de efectos fijos* (Modelo I). A un modelo lineal que únicamente incluye términos de efectos aleatorios se le llama *modelo de efectos aleatorios* (Modelo II). A un modelo lineal que incluye una mezcla de términos de efectos fijos y de efectos aleatorios se le llama *modelo de efectos mixtos* (Modelo III).

Gran parte de los procedimientos estadísticos disponibles se centran en el estudio de promedios y de la variación en torno a esos promedios. Un ejemplo significativo de este enfoque lo constituyen los modelos de análisis de varianza y covarianza estudiados en el capítulo anterior y en los Capítulos 6 al 9 del segundo volumen. Estos procedimientos coinciden en expresar una observación como el resultado de la combinación lineal de un conjunto de efectos; también coinciden en asumir que los datos se ajustan a una distribución normal (estos procedimientos y otros como el análisis de regresión lineal, que nosotros hemos agrupado bajo la denominación de *modelos lineales clásicos*, son expresiones concretas del llamado *modelo lineal general*).

Cuando los parámetros de un modelo lineal se interpretan como *constantes* fijas, los efectos asociados a esos parámetros también se consideran de efectos fijos. Los niveles concretos que adopta un factor de efectos fijos son todos los niveles que interesa estudiar (o todos los niveles que tiene el factor); por este motivo la hipótesis nula referida a un factor de efectos fijos se plantea justamente sobre las medias poblacionales correspondientes a los niveles del factor (y las inferencias se limitan a esos niveles concretos):

$$H_0: \mu_1 = \mu_2 = \dots = \mu_j = \dots = \mu_J \quad [3.1]$$

Una variante del modelo lineal general consiste en tratar los parámetros, no como constantes fijas, sino como *variables aleatorias*. Ya hemos explicado la diferencia existente entre efectos fijos y efectos aleatorios: los niveles concretos que toma un factor de efec-

tos aleatorios únicamente son una muestra aleatoria de la población de posibles niveles del factor. Por este motivo, la hipótesis nula referida a un factor de efectos aleatorios no se plantea sobre las medias de los niveles que toma el factor en un estudio concreto, sino sobre su *varianza*:

$$H_0: \sigma_{\beta}^2 = 0 \quad [3.2]$$

(σ_{β}^2 se refiere a la varianza poblacional de las medias de todos los posibles niveles del factor). Puesto que los J niveles de un factor de efectos aleatorios son solo algunos de los posibles, la hipótesis debe reflejar, no la igualdad entre las medias de esos J niveles, sino la igualdad entre todos los posibles niveles del factor. Tal como se afirma en [3.2], la varianza de esas medias valdrá cero cuando todas ellas sean iguales. Y dada la naturaleza de esta H_0 , las inferencias se realizarán, no sobre los J niveles incluidos en el análisis, sino sobre la población de niveles del factor. Por tanto, en un modelo de efectos aleatorios, el interés del análisis no se centra en las medias de los niveles del factor, sino en su varianza: lo que realmente interesa saber es en qué medida el término aleatorio contribuye a explicar la varianza de la variable dependiente.

Cuando un modelo lineal incluye una mezcla de efectos fijos y aleatorios, se tiene un modelo lineal de efectos mixtos o, simplemente, un **modelo lineal mixto**. Un modelo mixto no solo permite analizar promedios (objetivo primordial de los modelos de efectos fijos), sino la estructura de covarianza de los datos (objetivo primordial de los modelos de efectos aleatorios). Para más detalles, ver, en el Apéndice 3, el apartado *Elementos de un modelo lineal mixto*.

Modelos con grupos aleatorios

El procedimiento **MIXED** de SPSS (opción **Modelos mixtos > Lineales** del menú **Analizar**) permite ajustar modelos más generales que las opciones **Univariante** y **Medidas repetidas** (ver Capítulos 7 al 9 del segundo volumen) del procedimiento **GLM** (opción **Modelo lineal general** del menú **Analizar**).

En los modelos que permite ajustar el procedimiento **GLM** se establecen, además del ya mencionado de normalidad, otros dos supuestos básicos. En primer lugar, se asume que los errores (por tanto, las puntuaciones de la variable dependiente) son independientes entre sí y, en segundo lugar, que se distribuyen de idéntica forma en todos los grupos (es decir, que las varianzas poblacionales de todos los grupos son iguales).

Pues bien, la principal ventaja del procedimiento **MIXED** es que permite relajar tanto el supuesto de independencia (permite analizar observaciones relacionadas, lo cual resulta especialmente útil para ajustar modelos de medidas repetidas) como el de igualdad de varianzas (permite trabajar con diferentes estructuras de covarianza). Por tanto, el procedimiento **MIXED** sirve para abordar estructuras de datos complejas de una forma más flexible que el procedimiento **GLM**. Y también permite ajustar modelos *multinivel* (modelos que permiten analizar datos de una muestra seleccionada de subgrupos pertenecientes a grupos de mayor orden; ver siguiente capítulo).

El hecho de que el interés principal de la mayoría de los experimentos esté, de hecho, centrado en el análisis de los efectos fijos podría hacer pensar que el procedimiento **GLM** contiene todo lo necesario para efectuar un análisis correcto. Sin embargo, en el ámbito de las ciencias sociales y de la salud no es infrecuente tener que trabajar con observaciones relacionadas y varianzas heterogéneas. En estos casos, que el procedimiento **GLM** no permite tratar apropiadamente en toda su complejidad, el procedimiento **MIXED** permite efectuar las correcciones oportunas, no ya solo para estimar los efectos aleatorios, sino para trabajar correctamente con los efectos fijos.

Análisis de varianza: un factor de efectos aleatorios

Este apartado muestra cómo estimar e interpretar los elementos del modelo de ANOVA de un factor de efectos aleatorios. El ejemplo se basa en el archivo *Depresión* (puede descargarse de la página web del manual), el cual contiene una muestra de 379 pacientes afectados de trastorno depresivo que han recibido tratamiento en 11 centros hospitalarios distintos. La variable *recuperación* refleja la recuperación experimentada a las 6 semanas del tratamiento; se ha obtenido calculando la diferencia entre las puntuaciones del momento basal (*basal*) y las de la semana 6 (*sexta*) en la escala de depresión de Hamilton (las puntuaciones más altas en esta escala indican mayor depresión). La variable *centro* indica el centro hospitalario donde han sido tratados los pacientes.

¿Cómo saber si los centros difieren en el nivel de recuperación que alcanzan los pacientes al cabo de 6 semanas de tratamiento? Esta pregunta podría responderse comparando la recuperación media de los 11 centros mediante un ANOVA de un factor. Pero esta solución trata al factor *centro* como un efecto fijo y las inferencias que es posible hacer se limitan a los centros incluidos en el análisis. Es más apropiado tratar los centros como un efecto aleatorio. De hecho, el objetivo del estudio no es averiguar si existen diferencias entre los 11 centros incluidos en el análisis¹; lo que interesa averiguar es, más bien, en qué medida la recuperación observada puede ser atribuida a diferencias entre los centros.

Por otro lado, el modelo de efectos fijos asume que las observaciones son independientes entre sí. Sin embargo, la recuperación de los pacientes de un mismo centro (misma ambiente, mismo equipo médico, etc.) es muy posible que sea más parecida que la recuperación de pacientes de centros distintos. Todas estas razones recomiendan utilizar un modelo de efectos aleatorios.

En el ANOVA de un factor de efectos fijos, la variable dependiente Y_{ij} se interpreta como el resultado de combinar un término constante (μ), el efecto atribuible al factor (α_j) y el efecto atribuible a todo lo que no es ni la constante ni el factor, es decir,

¹ Dado que los 11 centros incluidos en el análisis constituyen una muestra aleatoria de la población de centros, carece de interés averiguar si tal centro concreto difiere de tal otro. El estudio de estas diferencias tendría sentido si el factor *centro* fuera de efectos fijos. Y sería de efectos fijos si, por ejemplo, interesando comparar la recuperación media de tres centros hospitalarios concretos de una determinada ciudad, se seleccionaran muestras aleatorias de esos tres centros.

los errores (E_{ij}). El modelo de efectos aleatorios incluye los mismos efectos que el de efectos fijos y, consecuentemente, su formulación es similar:

$$Y_{ij} = \mu + \beta_j + E_{ij} \quad [3.3]$$

(i se refiere a los casos: $i = 1, 2, \dots, n_j$; y j a los niveles del factor: $j = 1, 2, \dots, J$). Tanto en el modelo de efectos fijos como en el de efectos aleatorios se considera que el término μ es una constante, pero en el modelo de efectos fijos se interpreta como la media poblacional de los J niveles del factor incluidos en el análisis (la recuperación media obtenida en los 11 centros), mientras que en el modelo de efectos aleatorios se interpreta como la media poblacional de todos los posibles niveles del factor (de los cuales los 11 centros incluidos en el análisis solo son una muestra aleatoria).

En el modelo de efectos fijos se asume que los términos α_j son parámetros fijos, es decir, valores únicos y desconocidos de la población. En el modelo de efectos aleatorios se asume que los términos β_j son niveles de una variable aleatoria que se distribuye normalmente con media 0 y varianza σ_β^2 , e independientemente de los errores. En ambos modelos se asume que los errores son independientes entre sí y que se distribuyen normalmente con media 0 y varianza σ_E^2 . Por tanto, $\mathbf{R}$, es decir, la matriz de varianzas-covarianzas de los errores (ver Apéndice 3) es igual a $\sigma_E^2 \mathbf{I}$ (una matriz de tamaño $n \times n$, con σ_E^2 en la diagonal principal y ceros fuera de la diagonal).

Ya hemos señalado que, cuando un factor es de efectos *fijos*, los J niveles que adopta son todos los niveles que interesa estudiar (esos J niveles constituyen la población de niveles del factor); por este motivo la hipótesis nula se plantea justamente sobre las medias poblacionales de esos niveles. Por el contrario, cuando un factor es de efectos aleatorios, los niveles concretos que adopta únicamente constituyen una muestra aleatoria de la población posibles niveles; por este motivo la hipótesis nula no se plantea sobre las medias de los niveles, sino sobre su *varianza*. Ahora bien, como se está asumiendo que el factor es independiente de los errores, se verifica

$$\sigma_Y^2 = \sigma_\beta^2 + \sigma_E^2 \quad [3.4]$$

(puesto que el término μ es una constante, su varianza vale 0). En consecuencia, la varianza total de Y es la suma de dos componentes independientes: la varianza del factor y la varianza de los errores. De ahí el nombre de **componentes de la varianza** que suele darse a este modelo (para profundizar en los detalles de este modelo, puede consultarse Rao y Kleffe, 1988, o Searle, Casella y McCulloch, 1992).

Además de asumirse que el factor es independiente de los errores, cuando se trabaja con un factor de efectos aleatorios se está imponiendo una determinada estructura de covarianza a los datos: se está asumiendo que los niveles del factor son independientes entre sí y que la relación entre observaciones de un mismo nivel del factor es constante².

² Es decir, se está asumiendo, en primer lugar, que los pacientes de centros distintos se comportan de forma independiente; por tanto, $Cov(Y_{ij}, Y_{i'j'}) = 0$. Y, en segundo lugar, que la relación entre pacientes de un mismo centro es constante; en concreto, $Cov(Y_{ij}, Y_{i'j}) = Cov(\mu + \beta_j + E_{ij}, \mu + \beta_j + E_{i'j}) = Cov(\beta_j, \beta_j) = Var(\beta_j) = \sigma_\beta^2$.

En concreto, se está asumiendo que la matriz **G** de varianzas-covarianzas de los efectos aleatorios (ver Apéndice 3) es una matriz de tamaño $J \times J$, con σ_p^2 en la diagonal principal y ceros fuera de la diagonal (J se refiere al número de niveles del factor).

Para tratar el factor *centro* (archivo *Depresión*) como un factor de efectos aleatorios y obtener las estimaciones que ofrece el procedimiento **MIXED**:

- Seleccionar la opción **Modelos mixtos > Lineales** del menú **Analizar** para acceder al cuadro de diálogo *Modelos lineales mixtos: Especificar sujetos y medidas repetidas*. Este cuadro de diálogo, previo al principal, permite indicar qué variable o variables de las que posteriormente se utilizarán en el análisis sirven para identificar a los *sujetos* y qué variables representan *medidas repetidas* (más adelante veremos que este cuadro de diálogo también permite seleccionar una estructura de covarianza para las medidas repetidas³).
- En este cuadro de diálogo previo al principal, pulsar el botón **Continuar** (sin seleccionar ninguna variable) para acceder al cuadro de diálogo principal⁴.
- Seleccionar la variable *recuperación* (recuperación en la semana 6) y trasladarla al cuadro **Variable dependiente**; seleccionar la variable *centro* (centro hospitalario) y trasladarla a la lista **Factores**.
- Pulsar el botón **Aleatorios** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos aleatorios* y trasladar la variable *centro* a la lista **Modelo**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

³ En un archivo convencional es habitual que los sujetos (los casos) constituyan unidades de observación independientes entre sí. Pero esto no tiene por qué ser siempre así. En un estudio con pacientes de diferentes hospitales, la variable *hospital* agrupa a pacientes que se parecen entre sí (al menos en parte) y que difieren de los pacientes de otros hospitales (también en parte); en un estudio con alumnos de distintos colegios, la variable *colegio* agrupa a alumnos que se parecen entre sí y que difieren de los alumnos de otros colegios. Si se desea que los hospitales o los colegios definan unidades de observación independientes entre sí, es necesario trasladar estas variables a la lista **Sujetos** (debe tenerse en cuenta que este tipo de variables no siempre intervienen en un modelo de ANOVA).

La lista **Repetidas** permite indicar qué variables representan medidas repetidas. El procedimiento **MIXED** exige que las medidas repetidas estén dispuestas de una forma particular (ver, más adelante, el apartado *Modelos de medidas repetidas*). Y el menú desplegable **Tipo de covarianza para repetidas** permite seleccionar el tipo de estructura de covarianza que se desea asignar a la matriz de varianzas-covarianzas residual (**R**) en los diseños de medidas repetidas (ver, más adelante, el apartado *Estructura de la matriz de varianzas-covarianzas residual*).

⁴ La lista de variables muestra un listado con todas las variables del archivo de datos, incluidas las que tienen formato de cadena. El significado de las listas **Variable dependiente**, **Factores** y **Covariables** es el mismo que en otros cuadros de diálogo ya estudiados. La opción **Ponderación de los residuos** sirve para ajustar modelos en los que se incumple el supuesto de varianzas constantes. En un modelo lineal clásico se asume que la varianza de la variable dependiente es la misma en todas las poblaciones objeto de estudio (en un diseño factorial estas poblaciones son tantas como casillas resultan de la combinación de los niveles de los factores). Cuando las varianzas poblacionales no son iguales (como, por ejemplo, cuando las casillas con puntuaciones mayores tienen más variabilidad que las casillas con puntuaciones menores), los métodos de estimación no consiguen ofrecer estimaciones óptimas. En estos casos, si la variabilidad de las casillas se conoce o puede estimarse a partir de alguna variable, es posible tener en cuenta esa variabilidad al estimar los parámetros de un modelo lineal. Al seleccionar una variable de ponderación se da más importancia a las observaciones más precisas, es decir, a aquellas con menor variabilidad (un valor frecuentemente utilizado para ponderar los residuos es el valor inverso de la matriz de varianzas-covarianzas). La variable de ponderación debe ser cuantitativa y su formato numérico (el procedimiento no permite ponderar con variables de cadena). Los valores de la variable de ponderación se tratan de forma similar a como se hace con los pesos de la regresión lineal.

- Pulsar el botón **Estadísticos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Estadísticos* y marcar las opciones **Estadísticos descriptivos**, **Estimaciones de los parámetros** y **Contrastes sobre los parámetros de covarianza**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* de resultados ofrece la información que muestran las Tablas 3.1 a 3.7.

Información preliminar

La Tabla 3.1 contiene *información descriptiva*. El número de pacientes por centro oscila entre 15 y 82. La recuperación media observada no es la misma en todos los centros; en el centro nº 5 se obtiene la media más baja (4,50); en el nº 11, la más alta (13,40); a la espera de lo que puedan decir los contrastes pertinentes, la recuperación parece estar relacionada con el centro. Las últimas dos columnas ofrecen la desviación típica y el coeficiente de variación (cociente entre la desviación típica y la media, expresado en porcentaje).

Tabla 3.1. Estadísticos descriptivos

Recuperación en la semana 6				
Centro hospitalario	n	Media	Desviación típica	Coeficiente de variación
1	23	11,91	4,316	36,2%
2	15	7,80	4,329	55,5%
3	27	8,81	4,333	49,2%
4	35	6,31	3,886	61,5%
5	34	4,50	3,518	78,2%
6	17	11,24	4,684	41,7%
7	24	10,58	4,772	45,1%
8	32	7,09	4,496	63,4%
9	42	13,10	5,026	38,4%
10	48	5,85	3,632	62,0%
11	82	13,40	4,139	30,9%
Total	379	9,51	5,325	56,0%

La Tabla 3.2 resume la *dimensión del modelo* propuesto. El modelo incluye tres efectos: un efecto fijo (la constante o *intersección*) y dos efectos aleatorios (el factor y los errores o *residuos*). La intersección (μ) es el parámetro de efectos fijos; la varianza de los

Tabla 3.2. Dimensión del modelo

		Nº de niveles	Estructura de covarianza	Nº de parámetros
Efectos fijos	Intersección	1	Componentes de la varianza	1
Efectos aleatorios	centro	11		1
Residuos				1
Total		12		3

za del factor (σ_{β}^2) y la varianza de los errores (σ_{ϵ}^2) son los dos parámetros de efectos aleatorios (tres parámetros en total). La tabla también informa de la estructura de covarianza (matriz **G**) impuesta al factor de efectos aleatorios: *componentes de la varianza* (es la estructura de covarianza que el SPSS utiliza por defecto).

Ajuste global

La Tabla 3.3 ofrece varios *estadísticos de ajuste global* que indican el grado en que el modelo propuesto se aleja del ajuste perfecto. El primero de estos estadísticos es la *desvianza* ($-2LL$). El resto son modificaciones de la desvianza que penalizan su valor (incrementándolo) mediante, básicamente, alguna función del número de parámetros⁵. Una nota a pie de tabla recuerda que el ajuste del modelo a los datos es tanto mejor cuanto menor es el valor de estos estadísticos (no olvidemos que la desvianza es una medida de *desajuste*).

Estos estadísticos no tienen una interpretación directa, pero son muy útiles para comparar modelos alternativos cuando uno de ellos incluye todos los términos del otro más uno o varios términos adicionales. La diferencia entre las desvianzas de dos modelos distintos (uno subconjunto del otro) es la *razón de verosimilitudes* G^2 (ver ecuación [1.15]). Este estadístico se distribuye según ji-cuadrado con los grados de libertad resultantes de restar el número de parámetros de los dos modelos comparados. Por tanto, la diferencia entre las desvianzas de dos modelos distintos (uno subconjunto del otro) sirve para: (1) cuantificar la reducción del desajuste asociada a los efectos en que difieren ambos modelos y (2) valorar la significación estadística de esa reducción.

En nuestro ejemplo, el efecto del factor *centro* puede evaluarse comparando la desvianza del modelo que incluye ese efecto (modelo 1) y la del modelo que no lo incluye (modelo 0). La Tabla 3.3 muestra la desvianza del modelo que incluye la intersección y el factor *centro* ($-2LL_1 = 2.199,27$). La Tabla 3.4 ofrece la desvianza del modelo que

⁵ El segundo estadístico (*AIC*) es el *criterio de información de Akaike* (Akaike, 1974):

$$AIC = -2LL + 2d \quad [3.5]$$

El tercer estadístico (*AICC*) es el *criterio de información de Akaike corregido* (Hurvich y Tsai, 1989):

$$AICC = -2LL + [2dn/(n - d - 1)] \quad [3.6]$$

El cuarto estadístico (*CAIC*) es el *criterio de información de Akaike consistente* (Bozdogan, 1987):

$$CAIC = -2LL + d[\log_e(n) + 1] \quad [3.7]$$

Y el quinto estadístico (*BIC*) es el *criterio de información bayesiano* (Schwarz, 1978; ver también Raftery, 1995):

$$BIC = -2LL + d[\log_e(n)] \quad [3.8]$$

En estas ecuaciones, *LL* se refiere al logaritmo de la verosimilitud si se utiliza el método de estimación MV (máxima verosimilitud) y al logaritmo de la verosimilitud restringida si se utiliza el método de estimación MVR (máxima verosimilitud restringida). Cuando se utiliza MV, *d* se refiere al número de parámetros asociados a los efectos fijos más el número de parámetros asociados a los efectos aleatorios y *n* al número total de casos. Cuando se utiliza MVR, *d* se refiere al número de parámetros asociados a los efectos aleatorios y *n* al número total de casos menos el número de parámetros asociados a los efectos fijos.

Tabla 3.3. Estadísticos de ajuste global (modelo 1: incluye la intersección y el factor *centro*)

-2 log de la verosimilitud restringida	2199,27
Criterio de información de Akaike (AIC)	2203,27
Criterio de Hurvich y Tsai (AICC)	2203,30
Criterio de Bozdogan (CAIC)	2213,14
Criterio bayesiano de Schwarz (BIC)	2211,14

Los criterios de información se muestran en formatos de mejor cuanto más pequeños.

únicamente incluye la intersección ($-2LL_0 = 2.342,94$; para obtener esta tabla hay que ajustar un modelo sin variables independientes). La diferencia entre ambas desviaciones

$$G^2_{0-1} = -2LL_0 - (-2LL_1) = 2.342,94 - 2.199,27 = 143,67$$

se distribuye según ji-cuadrado con 1 grado de libertad (el correspondiente al parámetro asociado al factor *centro* –ver Tabla 3.2–, que es el único parámetro en el que difieren ambos modelos). La probabilidad de encontrar valores ji-cuadrado iguales o mayores que 143,67 es menor que 0,0005 (esta probabilidad puede obtenerse utilizando la expresión `SIG.CHISQ(143.67, 1)` en el cuadro de diálogo **Calcular** del menú **Transformar**); por tanto, se puede rechazar la hipótesis de que el efecto del factor *centro* es nulo.

Aunque la valoración de un efecto concreto forma parte de los resultados del procedimiento **MIXED**, con tamaños muestrales pequeños la estrategia basada en el cambio observado en el estadístico $-2LL$ es preferible a la basada en el estadístico de *Wald* (ver Tabla 3.7). No obstante, la verdadera utilidad de esta estrategia basada en el estadístico $-2LL$ radica en la posibilidad de comparar el efecto simultáneo de varios términos y, consecuentemente, en la posibilidad de comparar el ajuste de modelos rivales.

Tabla 3.4. Estadísticos de ajuste global (modelo 0: únicamente incluye la intersección)

-2 log de la verosimilitud restringida	2342,94
Criterio de información de Akaike (AIC)	2344,94
Criterio de Hurvich y Tsai (AICC)	2344,95
Criterio de Bozdogan (CAIC)	2349,87
Criterio bayesiano de Schwarz (BIC)	2348,87

Los criterios de información se muestran en formatos de mejor cuanto más pequeños.

Significación de los efectos incluidos en el modelo

La Tabla 3.5 ofrece los *contrastes de los efectos fijos* incluidos en el modelo. El modelo de un factor de efectos aleatorios únicamente contiene un efecto fijo: la constante o intersección. La intersección es una estimación de la *media poblacional* (es la media no ponderada de la recuperación media de los 11 centros; la Tabla 3.6 indica que esa media vale 9,15). Y el estadístico *F* que ofrece la tabla permite contrastar la hipótesis nula de que esa media poblacional vale cero. Puesto que el nivel crítico (*sig.* < 0,0005) aso-

ciado al estadístico F es menor que 0,05, se puede concluir que la recuperación media en la población de centros es mayor que cero. El contraste de la hipótesis nula referida a la intersección no suele tener interés, sin embargo, el rechazo de esta hipótesis en nuestro ejemplo está indicando que la recuperación media es mayor que cero.

Tabla 3.5. Contraste de los efectos fijos (sumas de cuadrados Tipo III)

Origen	Numerador df	Denominador df	Valor F	Sig.
Intersección	1	10,30	94,62	,000

Estimaciones de los parámetros

Las Tablas 3.6 y 3.7 ofrecen las estimaciones de los parámetros. La Tabla 3.6 recoge una *estimación de la constante o intersección* (el único parámetro de efectos fijos que incluye el modelo). El valor estimado (9,15) aparece acompañado de su error típico, de su valor tipificado ($9,15 / 0,94 = 9,73$), de los grados de libertad de su distribución muestral, del nivel crítico obtenido al contrastar la hipótesis nula de que la intersección vale cero en la población, y de los límites inferior y superior del intervalo de confianza calculado al 95%. Se considera que un parámetro es significativamente distinto de cero cuando su nivel crítico (*sig.*) es menor que 0,05; o, lo que es equivalente, cuando su intervalo de confianza no incluye el valor cero. Los resultados de la tabla permiten concluir que la intersección es distinta de cero (*sig.* < 0,0005).

Tabla 3.6. Estimaciones de los parámetros de efectos fijos

Parámetro	Estimación	Error típico	gl	t	Sig.	Intervalo de confianza 95%	
						L. inferior	L. superior
Intersección	9,15	,94	10,30	9,73	,000	7,06	11,23

Finalmente, la Tabla 3.7 ofrece las *estimaciones de los parámetros de covarianza*, es decir, de los parámetros asociados a los dos efectos aleatorios que incluye el modelo. La *varianza del factor* (*centro* = 9,09) es una estimación de la variabilidad existente entre las medias de los centros (σ_p^2). La *varianza de los residuos* (*residuos* = 18,00) es una estimación de la variabilidad error, es decir, de la variabilidad existente dentro de cada centro (σ_E^2); esta varianza es la misma que en otros procedimientos SPSS recibe el nombre de *media cuadrática error*.

Dividiendo la varianza del factor entre la suma de ambas varianzas se obtiene el **coeficiente de correlación intraclase**:

$$C_{CI} = \sigma_p^2 / (\sigma_p^2 + \sigma_E^2) = 9,09 / (9,09 + 18,00) = 0,34$$

Este valor indica que la variabilidad entre los niveles del factor (las diferencias en la recuperación media de los centros) representa el 34% de la variabilidad total (es decir,

de la variabilidad de la recuperación). El coeficiente de correlación intraclase es una cuantificación del grado de variabilidad existente entre los centros en comparación con la variabilidad existente entre los pacientes del mismo centro. Un valor de uno indica que toda la variabilidad se debe al factor, es decir, a la diferencia entre los centros (lo que solo ocurrirá cuando en todos los pacientes de un mismo centro se dé la misma recuperación y los centros tengan diferentes promedios). Un coeficiente de cero indica que el factor no contribuye en absoluto a explicar la variabilidad de la recuperación; es decir, que toda la variabilidad está explicada por las diferencias existentes dentro de cada centro (lo que solo ocurrirá cuando la recuperación media de todos los centros sea la misma). Por tanto, el valor del C_{CI} también representa el grado de relación existente entre los pacientes del mismo centro.

Las estimaciones de los parámetros de covarianza que ofrece la Tabla 3.7 aparecen acompañadas de la información necesaria para obtener la significación estadística de cada estimación. La hipótesis que interesa contrastar en el modelo de un factor es que el efecto del factor es nulo. Y recordemos que, puesto que se trata de un factor de efectos aleatorios, esta hipótesis adopta la forma:

$$H_0: \sigma_{\beta}^2 = 0 \quad [3.9]$$

Para contrastar esta hipótesis, el SPSS ofrece el estadístico de *Wald* y un intervalo de confianza. El estadístico de Wald se obtiene dividiendo el correspondiente valor estimado entre su error típico: $9,09/4,28 = 2,12$ (la distribución muestral de este cociente se aproxima a la normal). Puesto que el nivel crítico asociado a 2,12 (*sig.* = 0,034) es menor que 0,05, se puede rechazar la hipótesis nula [3.9] y concluir que la varianza poblacional del factor es distinta de cero. Es decir, se puede concluir que la recuperación media no es la misma en todos los centros. El intervalo de confianza permite llegar a la misma conclusión, pues sus límites no incluyen el valor cero. Es importante recordar que esta conclusión no se refiere a los once centros incluidos en el análisis, sino a la población de centros de la cual estos once centros constituyen una muestra aleatoria.

Los parámetros de covarianza se han estimado asumiendo que el factor *centro* es independiente de los errores (*componentes de la varianza*), de ahí que a este modelo se le llame **modelo incondicional**: la varianza de los centros es distinta de cero independientemente de cualquier otra consideración. A este modelo también se le suele llamar **modelo nulo** pues, según veremos, en algunos contextos se utiliza como referente para contrastar, por comparación con él, la significación de otros términos (no confundir este modelo con el que únicamente incluye la intersección, que también se utiliza como referente con el que comparar otros modelos).

Tabla 3.7. Estimaciones de los parámetros de covarianza

Parámetro	Estimación	Error típico	Wald Z	Sig.	Intervalo de confianza 95%	
					L. inferior	L. superior
Residuos	18,00	1,33	13,57	,000	15,58	20,80
centro Varianza	9,09	4,28	2,12	,034	3,61	22,89

Análisis de varianza: dos factores de efectos mixtos

Los resultados del ejemplo anterior indican que el factor *centro* consigue explicar aproximadamente un tercio de la varianza de la *recuperación* (recordemos que la variabilidad entre los centros representaba un 34% de la variabilidad total). Una variable que podría contribuir a explicar parte de los dos tercios de la variabilidad todavía no explicada es el tipo de *tratamiento* aplicado (*tto*). Cada paciente del archivo *Depresión* ha recibido uno de tres tratamientos distintos. La Tabla 3.8 muestra el número de pacientes sometidos a cada tratamiento en cada centro. El tratamiento *estándar* se ha aplicado a 111 pacientes y el *combinado* a 214; los 54 pacientes restantes han recibido un tratamiento distinto de los dos anteriores (*otro*). En total, $n = 379$ pacientes.

Tabla 3.8. Número de pacientes por *centro* y *tratamiento*

Recuento		Centro hospitalario											Total
		1	2	3	4	5	6	7	8	9	10	11	
Tratamiento	Estándar	7	5	8	10	10	5	7	10	13	13	23	111
	Combinado	13	8	15	20	19	9	13	17	24	28	48	214
	Otro	3	2	4	5	5	3	4	5	5	7	11	54
Total		23	15	27	35	34	17	24	32	42	48	82	379

El factor *tratamiento* es de efectos *fijos* (interesa estudiar justamente los tratamientos incluidos en el análisis). El factor *centro* ya ha quedado dicho que es de efectos *aleatorios*. Por tanto, un modelo que incluye el efecto de ambos factores es un modelo de efectos *mixtos*:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + E_{ijk} \quad [3.10]$$

(i se refiere a los casos o puntuaciones individuales: $i = 1, 2, \dots, n_{jk}$; j se refiere a los niveles del factor de efectos fijos: $j = 1, 2, \dots, J$; y k se refiere a los niveles del factor de efectos aleatorios: $k = 1, 2, \dots, K$). El término constante μ sigue siendo, al igual que en el modelo de un factor, la media poblacional de la variable dependiente (la recuperación media en el conjunto total de centros). El efecto del factor *tratamiento* (el término α_j) es fijo, es decir, cada α_j es un valor único y desconocido de la población. El efecto del factor *centro* (el término β_k) es una variable aleatoria que se asume que se distribuye normalmente con media 0, varianza σ_β^2 e independientemente de los errores. El efecto de la interacción entre ambos factores, $(\alpha\beta)_{jk}$, es una variable aleatoria⁶ que se asume que se distribuye normalmente con media 0, varianza $\sigma_{\alpha\beta}^2$ e independien-

⁶ Recuérdese que un término que incluye simultáneamente efectos fijos y efectos aleatorios se considera un término de efectos aleatorios. Dicho de otra forma: un término compuesto se considera de efectos fijos únicamente si todos los términos simples que incluye son de efectos fijos.

temente de los errores y del término β_k . Y los errores E_{ijk} se asume que son independientes entre sí y del resto de términos del modelo, y que se distribuyen normalmente con media 0 y varianza constante σ_E^2 . Por tanto, $\mathbf{R} = \sigma_E^2 \mathbf{I}$; es decir, la matriz de varianzas-covarianzas residual $\mathbf{R}$ (ver Apéndice 3) es una matriz de tamaño $n \times n$, con σ_E^2 en la diagonal principal y ceros fuera de la diagonal. Puesto que se está asumiendo que los términos incluidos en el modelo son independientes entre sí, se verifica:

$$\sigma_Y^2 = \sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma_E^2 \quad [3.11]$$

(μ es una constante y, por tanto, su varianza vale 0; y lo mismo vale decir del término α_j en cada j). En consecuencia, la varianza total es la suma de tres componentes independientes (tres *componentes de la varianza*): la varianza del factor de efectos aleatorios, la varianza de la interacción entre los dos factores y la varianza de los errores.

Además, puesto que se está asumiendo que los niveles del factor de efectos aleatorios son independientes entre sí y que la relación entre observaciones de un mismo nivel del factor es constante, la matriz $\mathbf{G}$ (es decir la matriz de varianzas-covarianzas de los efectos aleatorios) es una matriz diagonal de tamaño $(K+JK)(K+JK)$, con σ_β^2 en la diagonal principal de las K primeras filas, $\sigma_{\alpha\beta}^2$ en la diagonal principal de las restantes JK filas (J se refiere al número de niveles del factor de efectos fijos y K al número de niveles del factor de efectos aleatorios), y ceros fuera de la diagonal principal.

Veamos con un ejemplo concreto cómo ajustar un modelo de efectos mixtos y cómo interpretar las estimaciones que ofrece el procedimiento **MIXED** (seguimos utilizando el archivo *Depresión*, el cual puede descargarse de la página web del manual):

- En el cuadro de diálogo previo al principal, pulsar el botón **Continuar** (sin seleccionar ninguna variable) para acceder al cuadro de diálogo principal.
- Seleccionar la variable *recuperación* (recuperación en la semana 6) y trasladarla al cuadro **Variable dependiente**; seleccionar las variables *tto* (tratamiento) y *centro* (centro hospitalario) y trasladarlas a la lista **Factores**.
- Pulsar el botón **Fijos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos fijos* y trasladar la variable *tto* a la lista **Modelo**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Aleatorios** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos aleatorios* y trasladar la variable *centro* a la lista **Modelo**. Seleccionar las variables *tto* y *centro* activando la opción **Interacción** en el menú desplegable y pulsar el botón **Añadir** para trasladar a la lista **Modelo** la interacción *tto* $\times$ *centro*. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Estadísticos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Estadísticos* y marcar las opciones **Estimaciones de los parámetros** y **Contrastes sobre los parámetros de covarianza**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Medias marginales estimadas** para acceder al cuadro de diálogo *Modelos lineales mixtos: Medias marginales estimadas* y trasladar la variable *tto* a la

lista **Mostrar las medias para**. Marcar la opción **Comparar los efectos principales** y, en el menú desplegable **Ajuste del intervalo de confianza**, seleccionar **Bonferroni**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones se obtienen, entre otros, los resultados que muestran las Tablas 3.9 a 3.15.

Información preliminar

La Tabla 3.9 comienza informando de los efectos que incluye el modelo: dos efectos fijos (la *intersección* y el factor *tto*) y dos efectos aleatorios (el factor *centro* y la interacción *tto* $\times$ *centro*) más el término residual. A continuación ofrece el número de niveles de cada efecto: para los efectos fijos, la intersección y los 3 tratamientos; para los aleatorios, los 44 niveles resultantes de sumar a los 11 centros las 33 combinaciones entre los 3 tratamientos y los 11 centros. La penúltima columna informa del tipo de estructura de covarianza que se está asumiendo para los efectos aleatorios: *componentes de la varianza* (es la estructura de covarianza que el procedimiento aplica por defecto). La última columna contiene el número de parámetros independientes o no redundantes de que consta el modelo (seis en total): la intersección (μ), los dos correspondientes a los niveles del factor *tto* (α_1 y α_2 ; α_3 es redundante), la varianza del factor *centro* (σ_B^2), la varianza de la interacción *tto* $\times$ *centro* ($\sigma_{\alpha\beta}^2$) y la varianza de los errores o residuos (σ_E^2).

Tabla 3.9. Dimensión del modelo

		Nº de niveles	Estructura de covarianza	Nº de parámetros
Efectos fijos	Intersección	1	Componentes de la varianza	1
	tto	3		2
Efectos aleatorios	centro + tto * centro	44		2
Residuos				1
Total		48		6

Ajuste global

La Tabla 3.10 muestra los *estadísticos de ajuste global*. La desvianza del modelo propuesto, es decir, la desvianza del modelo que incluye la intersección, el factor fijo *tto*, el factor aleatorio *centro* y la interacción *tto* $\times$ *centro* (modelo 1) vale 2.121,93. Recordemos que la desvianza del modelo que únicamente incluye la intersección (modelo 0) vale 2.342,94 (ver Tabla 3.4), y que la desvianza del modelo que incluye la intersección y el factor *centro* vale 2.199,27 (ver Tabla 3.3). La *razón de verosimilitudes*

$$G_{0-1}^2 = 2.342,94 - 2.121,93 = 221,01$$

es la cantidad en que el modelo mixto propuesto consigue reducir la desvianza del modelo que solo incluye la intersección. Esta diferencia se distribuye según el modelo de

probabilidad ji-cuadrado con 4 grados de libertad (la diferencia en el número de parámetros independientes de ambos modelos). En la distribución ji-cuadrado con 4 grados de libertad, la probabilidad de obtener valores mayores que 221,01 es menor que 0,0005, por lo que puede afirmarse que los efectos incluidos en el modelo mixto contribuyen a mejorar significativamente el ajuste.

Respecto del modelo que solo incluye el factor *centro*, el modelo mixto consigue reducir la desviación en $2.199,27 - 2.121,93 = 77,34$ puntos. La probabilidad de obtener valores ji-cuadrado mayores que 77,34 con 3 grados de libertad (número de parámetros independientes en que difieren ambos modelos) es menor que 0,0005. Por tanto, también puede afirmarse que los efectos extra que incluye el modelo propuesto (los *tratamientos* y la interacción entre los *tratamientos* y los *centros*) contribuyen a reducir significativamente el desajuste del modelo que incluye la intersección y el factor *centro*.

Tabla 3.10. Estadísticos de ajuste global

-2 log de la verosimilitud restringida	2121,93
Criterio de información de Akaike (AIC)	2127,93
Criterio de Hurvich y Tsai (AICC)	2128,00
Criterio de Bozdogan (CAIC)	2142,72
Criterio bayesiano de Schwarz (BIC)	2139,72

Los criterios de información se muestran en formatos de mejor cuanto más pequeños.

Significación de los efectos incluidos en el modelo

La Tabla 3.11 ofrece los *contrastes de los efectos fijos*. El modelo mixto que estamos ajustando incluye dos efectos fijos: la constante (*intersección*) y el factor de efectos fijos (*tto*). La tabla ofrece los estadísticos *F* necesarios para contrastar las hipótesis de que estos efectos son nulos. La hipótesis nula referida a la intersección afirma que su valor poblacional es cero; puesto que el nivel crítico (*sig.*) asociado al estadístico *F* es menor que 0,05, se puede rechazar esa hipótesis y concluir que el valor poblacional de la intersección es distinto de cero. La hipótesis nula referida al factor *tto* afirma que el efecto del factor es nulo, es decir, que la recuperación media es la misma con los tres tratamientos. El nivel crítico (*sig.* < 0,0005) permite rechazar esa hipótesis y concluir que la recuperación media no es la misma con los tres tratamientos; o, lo que es equivalente, que la recuperación está relacionada con los tratamientos. Puesto que los centros constituyen un factor de efectos aleatorios, la conclusión a la que hemos llegado (que la recuperación está relacionada con los tratamientos) se refiere no solo a los centros incluidos en el análisis sino a toda la población de centros.

Tabla 3.11. Contraste de los efectos fijos (sumas de cuadrados Tipo III)

Origen	Numerador df	Denominador df	Valor F	Sig.
Intersección	1	10,49	80,50	,000
tto	2	27,88	30,74	,000

Estimaciones de los parámetros

Las Tablas 3.12 y 3.13 contienen las estimaciones de los parámetros del modelo. La Tabla 3.12 ofrece las *estimaciones de los parámetros asociados a los efectos fijos*. El procedimiento fija en cero la última categoría o nivel del factor (esta circunstancia se indica en una nota a pie de tabla) y estima los parámetros correspondientes al resto de categorías por comparación con la que se ha fijado en cero. De los tres parámetros asociados a la variable *tratamiento*, el último de ellos (el correspondiente a la categoría 3 = “otro”) se ha fijado en cero y únicamente se han estimado los parámetros correspondientes a las categorías 1 = “estándar” y 2 = “combinado”. El valor de la *intersección* (7,56) es la media de la categoría que se ha fijado en cero: *tto* = 3 = “otro”. El valor estimado para la categoría *tto* = 1 = “estándar” es la diferencia entre las medias de las categorías *estándar* y *otro*: $6,85 - 7,56 = -0,71$ (ver Tabla 3.14). Y el valor estimado para la categoría *tto* = 2 = “combinado” es la diferencia entre las medias de las categorías *combinado* y *otro*: $10,79 - 7,56 = 3,23$.

La tabla incluye, para cada estimación, su error típico, sus grados de libertad, su valor tipificado *t* (cociente entre el valor estimado y su error típico), el nivel crítico obtenido al contrastar la hipótesis de que el correspondiente parámetro vale cero y el intervalo de confianza calculado al 95%. Se considera que un parámetro es distinto de cero cuando el correspondiente nivel crítico (*sig.*) es menor que 0,05; o, lo que es equivalente, cuando su intervalo de confianza al 95% no incluye el valor cero.

Los resultados de nuestro ejemplo indican que la diferencia entre los tratamientos *estándar* y *otro* no es significativa (*sig.* = 0,313) y que los tratamientos *combinado* y *otro* difieren significativamente (*sig.* < 0,0005). No obstante, esta no es la mejor manera de comparar los tratamientos pues, además de que falta una comparación (la correspondiente a los tratamientos *estándar* y *combinado*), no se está aplicando ninguna estrategia para controlar la tasa de error. Para realizar estas comparaciones es preferible utilizar la opción **Comparar los efectos principales** del subcuadro de diálogo *Medias marginales estimadas* (ver, más adelante, las comparaciones por pares que ofrece la Tabla 3.15).

Tabla 3.12. Estimaciones de los parámetros de efectos fijos

Parámetro	Estimación	Error típico	gl	t	Sig.	Intervalo de confianza 95%	
						L. inferior	L. superior
Intersección	7,56	1,06	17,22	7,13	,000	5,33	9,80
[tto=1]	-,71	,70	51,59	-1,02	,313	-2,12	,69
[tto=2]	3,23	,65	40,32	4,94	,000	1,91	4,55
[tto=3]	,00 ^a	,00	.	.	.	.	.

a. Se ha establecido este parámetro en cero porque es redundante.

La Tabla 3.13 muestra las *estimaciones de los parámetros de covarianza*. A estas estimaciones se les suele llamar *condicionadas* porque dependen de los efectos fijos presentes en el modelo. El modelo incluye tres parámetros de covarianza:

- La varianza de los *residuos* (σ_e^2) refleja la variabilidad de la recuperación dentro de cada centro; se trata de la variabilidad intracentro que todavía falta por explicar

después de incluir en el modelo el factor *tratamiento*, el factor *centro* y la interacción entre ambos; de los tres componentes de la varianza, éste es el mayor, pero se ha reducido en un 20% respecto del valor obtenido con el modelo que únicamente incluía el factor *centro* (ha bajado de 18,00 a 14,53; ver Tabla 3.7).

- La varianza del factor *centro* (σ_p^2) refleja la variabilidad entre las medias de los centros; su valor es similar al obtenido con el modelo de un factor (8,84 frente a 9,09; ver Tabla 3.7) y sigue siendo significativamente distinto de cero (*sig.* = 0,036).
- La varianza asociada al efecto de la interacción *tto* × *centro* ($\sigma_{\alpha\beta}^2$) no difiere significativamente de cero (*sig.* = 0,361). Por tanto, no parece que el efecto de los tratamientos cambie de un centro a otro, lo cual sugiere que la interacción *tto* × *centro* podría ser eliminada del modelo sin pérdida de ajuste.

Tabla 3.13. Estimaciones de los parámetros de covarianza

Parámetro	Estimación	Error típico	Wald Z	Sig.	Intervalo de confianza 95%	
					L. inferior	L. superior
Residuos	14,53	1,09	13,36	,000	12,54	16,82
centro Varianza	8,84	4,22	2,10	,036	3,47	22,51
tto * centro Varianza	,42	,46	,91	,361	,05	3,55

Comparaciones múltiples

Por último, los resultados incluyen las *medias estimadas* y las *comparaciones por pares* entre ellas.

Las medias estimadas que ofrece la Tabla 3.14 son las medias marginales no ponderadas. La tabla ofrece, para cada media estimada, el error típico, los grados de libertad y los límites del intervalo de confianza individual calculado al 95%.

Tabla 3.14. Medias marginales estimadas

Tratamiento	Media	Error típico	gl	Intervalo de confianza 95%	
				L. inferior	L. superior
Estándar	6,85	,99	13,22	4,71	8,99
Combinado	10,79	,96	11,58	8,69	12,89
Otro	7,56	1,06	17,22	5,33	9,80

Una vez estimadas las medias, el procedimiento las compara por pares para determinar cuáles de ellas difieren entre sí (ver Tabla 3.15). Estas comparaciones son idénticas a las comparaciones *post hoc* ya estudiadas en los Capítulos 6 al 9 del segundo volumen y se interpretan de la misma manera (el subcuadro de diálogo *Modelos lineales mixtos: Medias marginales estimadas* también contiene opciones para comparar, no cada media con cada otra, sino cada media con otra cualquiera, a elegir).

Los resultados de la Tabla 3.15 indican que la recuperación que se alcanza con el tratamiento *combinado* difiere significativamente de la que se alcanza con los otros dos tratamientos (*sig.* < 0,0005 en ambos casos); en concreto, la recuperación media es *más alta* con el tratamiento *combinado*. Y no existe evidencia de que la recuperación que se alcanza con el tratamiento *estándar* sea distinta de la que se alcanza con el tratamiento *otro* (*sig.* = 0,940).

Tabla 3.15. Comparaciones por pares entre las medias estimadas

(I) Tratamiento	(J) Tratamiento	Diferencia entre las medias (I-J)	Error típico	gl	Sig. ^a	Intervalo de confianza 95% para la diferencia ^a	
						L. inferior	L. superior
Estándar	Combinado	-3,94	,54	19,57	,000	-5,34	-2,54
	Otro	-,71	,70	51,59	,940	-2,44	1,02
Combinado	Otro	3,23	,65	40,32	,000	1,60	4,86

Basado en las medias marginales estimadas

a. Corrección por comparaciones múltiples: Bonferroni.

Modelos con medidas repetidas

En los Capítulos 8 y 9 del segundo volumen hemos estudiado los modelos de medidas repetidas tal como permite abordarlos la opción **Modelo lineal general > Medidas repetidas** del SPSS (procedimiento **GLM**). Esta forma de analizar los datos impone algunas restricciones.

En primer lugar, dado que cada medida repetida se registra en el archivo de datos como una variable (una columna), el procedimiento asume que las mediciones se han llevado a cabo en el mismo momento o a intervalos temporales idénticos. En segundo lugar, únicamente se consideran casos válidos para el análisis los que no tienen ningún valor perdido; es decir, los casos con algún valor perdido son excluidos del análisis. En tercer lugar, el procedimiento no permite definir modelos personalizados que incluyan solo algunas de las interacciones posibles; es posible definir interacciones entre factores intrasujetos; también es posible definir interacciones entre factores intersujetos; pero no es posible definir interacciones entre factores inter e intrasujetos (o se incluyen todas o ninguna). Por último, los estadísticos *F* univariados que ofrece el procedimiento asumen que la matriz de varianzas-covarianzas es esférica; aunque es posible aplicar algunas correcciones a los estadísticos univariados cuando se incumple este supuesto, no es posible elegir diferentes estructuras de covarianza.

Analizar medidas repetidas con el procedimiento **MIXED** posee algunas ventajas. En primer lugar, no exige que las medidas estén igualmente espaciadas. En segundo lugar, los casos con algún valor perdido pueden ser incluidos en el análisis (estas dos primeras consideraciones son de especial importancia si se tiene en cuenta que tanto en los ensayos clínicos como en los estudios experimentales es frecuente encontrar que el tiempo transcurrido entre medidas no suele ser exactamente el mismo en todos los sujetos y que

a algunos de ellos les falta alguna medida). En tercer lugar, es posible definir exactamente las interacciones que interesa estudiar. Por último, es posible elegir, entre distintas estructuras de covarianza, la que mejor se ajuste a los datos.

En este apartado se explica cómo utilizar el procedimiento **MIXED** para ajustar los mismos modelos de ANOVA que hemos ajustado en los Capítulos 8 y 9 del segundo volumen con el procedimiento **GLM**. Ahora bien, para analizar medidas repetidas con el procedimiento **MIXED** hay que tener en cuenta que la disposición que deben adoptar los datos en el *Editor de datos* difiere de la descrita a propósito del procedimiento **GLM**. Un par de ejemplos ayudarán a entender esto.

Estructura de los datos

La Tabla 3.16 muestra unos datos ya analizados en el Capítulo 8 del segundo volumen. Se han obtenido de un estudio sobre el efecto del *paso del tiempo* en la *calidad del recuerdo*. El diseño incluye 6 sujetos a los que se les ha hecho memorizar una historia cuyo recuerdo ha sido evaluado al cabo de una *hora*, un *día*, una *semana* y un *mes*. Se trata, por tanto, de un diseño de un factor (al que llamaremos *tiempo*) con cuatro niveles (los cuatro momentos en los que se registra el recuerdo: al cabo de una *hora*, un *día*, una *semana* y un *mes*) y una variable dependiente (la *calidad del recuerdo*; las puntuaciones más altas indican mejor recuerdo).

El procedimiento **GLM** (ver Capítulo 8 del segundo volumen) requiere que cada nivel del factor (*hora*, *día*, *semana* y *mes*) esté registrado en el archivo de datos como una variable distinta. El procedimiento **MIXED** requiere organizar los datos de otra manera. Puesto que el diseño únicamente incluye dos variables (la variable independiente o factor *tiempo* y la variable dependiente o respuesta *recuerdo*), el archivo de datos solo necesita incluir estas dos variables (al margen de la identificación de los casos).

La Figura 3.1 muestra cómo organizar los datos de la Tabla 3.16 para poder aplicar el procedimiento **MIXED**. Se trata de una reproducción parcial: la figura solo muestra los 2 primeros sujetos (*id*), los cuales ocupan las primeras 8 filas; cada sujeto ocupa 4 filas; el archivo con los 6 sujetos del ejemplo tiene 24 filas.

Tabla 3.16. Calidad del *recuerdo* al cabo del *tiempo*

<i>Sujetos</i>	<i>Hora</i>	<i>Día</i>	<i>Semana</i>	<i>Mes</i>	<i>Medias</i>
1	16	11	9	8	11
2	14	8	4	2	7
3	19	13	7	9	12
4	17	10	8	9	11
5	16	14	8	6	11
6	20	16	12	8	14
<i>Medias</i>	17	12	8	7	11

Figura 3.1. Datos de la Tabla 3.16

	<i>id</i>	<i>tiempo</i>	<i>recuerdo</i>
1	1	1	16
2	1	2	11
3	1	3	9
4	1	4	8
5	2	1	14
6	2	2	8
7	2	3	4
8	2	4	2

Los datos de la Tabla 3.17 se han analizado ya en el Capítulo 9 del segundo volumen. A una muestra aleatoria de 6 sujetos se les ha hecho memorizar dos listas distintas: una de *letras* y otra de *números*. Más tarde, al cabo de una *hora*, un *día*, una *semana* y un *mes*, se les ha solicitado reproducir ambas listas y, como una medida de la calidad del recuerdo, se ha contabilizado el número de aciertos. La Tabla 3.17 muestra los resultados obtenidos. Se trata de un diseño con dos factores, ambos con medidas repetidas. El primer factor, *contenido*, tiene 2 niveles: *números* y *letras*. El segundo factor, *tiempo*, tiene 4 niveles: *hora*, *día*, *semana* y *mes*. La Figura 9.1 del segundo volumen muestra cómo organizar los datos para utilizar la opción **Medidas repetidas** del procedimiento **GLM**. La forma de organizar los datos para utilizar el procedimiento **MIXED** es distinta. Puesto que el diseño consta de tres variables (dos variables independientes o factores –*tiempo* y *contenido*– y una variable dependiente o respuesta –*recuerdo*–), el archivo de datos únicamente necesita incluir estas tres variables.

La Figura 3.2 muestra cómo reproducir los datos de la Tabla 3.17 en el *Editor de datos* del SPSS. Cada sujeto ocupa 8 filas. La tabla únicamente muestra los 2 primeros sujetos, es decir, 16 filas; el archivo con los 6 sujetos del ejemplo tiene 48 filas. Los códigos 1 y 2 asignados al factor *contenido* corresponden a los niveles *números* y *letras*, respectivamente; los códigos 1, 2, 3 y 4 asignados al factor *tiempo* corresponden a una *hora*, un *día*, una *semana* y un *mes*, respectivamente.

Tabla 3.17. Recuerdo de números y letras al cabo del tiempo

<i>Sujetos</i>	<i>Números</i>				<i>Letras</i>			
	<i>Hora</i>	<i>Día</i>	<i>Semana</i>	<i>Mes</i>	<i>Hora</i>	<i>Día</i>	<i>Semana</i>	<i>Mes</i>
1	6	6	3	2	8	6	4	3
2	7	5	5	5	10	8	5	2
3	4	2	1	3	7	7	2	2
4	7	5	3	4	11	9	3	6
5	6	4	4	5	10	6	4	3
6	5	2	1	1	9	4	3	5

Figura 3.2 Datos de la Tabla 3.17 reproducidos en el *Editor de datos* (izqda.: caso nº 1; dcha.: caso nº 2)

	id	contenido	tiempo	recuerdo		id	contenido	tiempo	recuerdo
1	1	1	1	6	9	2	1	1	7
2	1	1	2	6	10	2	1	2	5
3	1	1	3	3	11	2	1	3	5
4	1	1	4	2	12	2	1	4	5
5	1	2	1	8	13	2	2	1	10
6	1	2	2	6	14	2	2	2	8
7	1	2	3	4	15	2	2	3	5
8	1	2	4	3	16	2	2	4	2

La diferencia fundamental en la disposición de los datos cuando se utilizan los procedimientos **GLM** y **MIXED** está en el número de filas que ocupa cada sujeto en el archivo de datos. Para utilizar el procedimiento **GLM**, cada sujeto debe ocupar una fila; para utilizar el procedimiento **MIXED**, cada sujeto debe ocupar tantas filas como medidas repetidas tenga el diseño; es decir, cada valor de la variable dependiente debe ocupar una fila.

Análisis de varianza: un factor con medidas repetidas

En la notación propia de los modelos de ANOVA, el modelo de un factor de medidas repetidas adopta la forma:

$$Y_{ij} = \mu + \alpha_j + S_i + E_{ij} \quad [3.12]$$

donde μ es la media poblacional de la variable dependiente, α_j es el efecto del factor (las diferencias entre las medias de las medidas repetidas) y S_i es la variabilidad entre las medias de los sujetos. Los E_{ij} siguen siendo los errores aleatorios. El modelo asume que S_i y E_{ij} son variables aleatorias independientes del resto de términos del modelo e independientes entre sí, y distribuidas normalmente con varianzas σ_S^2 y σ_E^2 , respectivamente. Puesto que tanto μ como α_j tienen varianza nula en cada una de las J poblaciones del diseño (pues μ es una constante y α_j es constante en cada j), se verifica:

$$\sigma_Y^2 = \sigma_S^2 + \sigma_E^2 \quad [3.13]$$

Por tanto, al igual que ocurre en el modelo de un factor de efectos aleatorios, en el modelo de un factor de medidas repetidas se verifica que la variabilidad total es la suma de dos componentes independientes (*componentes de la varianza*): la varianza de los sujetos (variabilidad intersujetos) y la varianza de los errores (variabilidad intrasujetos). En el Capítulo 8 del segundo volumen se ofrece una descripción de las características de este modelo y de los efectos que interesa analizar.

Para ajustar un modelo de medidas repetidas a los datos de la Tabla 3.16 con el procedimiento **MIXED** (los datos se encuentran en el archivo *Tiempo recuerdo*, el cual puede descargarse de la página web del manual):

- En el cuadro de diálogo previo al principal⁷, trasladar la variable *id* (identificación de caso) a la lista **Sujetos** y la variable *tiempo* a la lista **Repetidas**; seleccionar **Simetría compuesta** en el menú desplegable **Tipo de covarianza para repetidas** y pulsar el botón **Continuar** para acceder al cuadro de diálogo principal.

⁷ Acabamos de ver que el procedimiento **MIXED** exige que las medidas repetidas estén dispuestas de una forma particular. La lista **Sujetos** sirve para indicar qué variable del archivo identifica a cada sujeto. La lista **Repetidas** sirve para indicar qué variable del archivo identifica a las medidas repetidas. El menú desplegable **Tipo de covarianza para repetidas** permite seleccionar un tipo de estructura de covarianza para la matriz de varianzas-covarianzas residual (**R**) en los diseños de medidas repetidas (ver, más adelante, el apartado *Estructura de la matriz de varianzas-covarianzas residual*).

- Trasladar la variable *recuerdo* (calidad del recuerdo) al cuadro **Variable dependiente** y la variable *tiempo* a la lista **Factores**.
- Pulsar el botón **Fijos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos fijos* y trasladar la variable *tiempo* a la lista **Modelo**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Estadísticos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Estadísticos* y marcar las opciones **Estimaciones de los parámetros**, **Contrastes sobre los parámetros de covarianza** y **Covarianzas de los residuos**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Medias marginales estimadas** para acceder al cuadro de diálogo *Modelos lineales mixtos: Medias marginales estimadas* y trasladar la variable *tiempo* a la lista **Mostrar las medias para**. Marcar la opción **Comparar los efectos principales** y, en el menú desplegable **Corrección del intervalo de confianza**, seleccionar **Bonferroni** (esta es la forma de solicitar comparaciones *post hoc* entre los niveles de un factor intrasujetos). Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor* ofrece, entre otros, los resultados que muestran las Tablas 3.18 a 3.23.

Significación de los efectos incluidos en el modelo

La Tabla 3.18 ofrece los *contrastes de los efectos fijos*. El modelo que estamos ajustando incluye dos efectos fijos: la constante o *intersección* y el factor *tiempo*. Los estadísticos *F* que ofrece la tabla permiten contrastar las hipótesis de que ambos efectos son nulos (estos estadísticos *F* son idénticos a los que se obtienen con la opción **Medidas repetidas** del procedimiento **GLM** (esfericidad asumida).

La *intersección* es la media de la variable dependiente (calidad del recuerdo) al cabo de un mes (momento que el procedimiento fija en cero; ver Tabla 3.19) y la hipótesis nula afirma que esa media vale cero. Puesto que el valor del correspondiente nivel crítico es muy pequeño (*sig.* < 0,0005), se puede rechazar esa hipótesis y concluir que la calidad del recuerdo al cabo de un mes es distinta de cero.

La hipótesis nula referida al factor *tiempo* afirma que el efecto del factor es nulo, es decir, que la calidad del recuerdo es la misma en los cuatro momentos. El valor del nivel crítico permite rechazar esa hipótesis nula (*sig.* < 0,0005) y concluir que la calidad del recuerdo no es la misma en los cuatro momentos incluidos en el análisis; o, lo que es lo mismo, que la calidad del recuerdo está relacionada con el paso del tiempo.

Tabla 3.18. Contraste de los efectos fijos (sumas de cuadrados Tipo III)

Origen	Numerador df	Denominador df	Valor F	Sig.
Intersección	1	5	139,62	,000
tiempo	3	15	58,13	,000

Estimaciones de los parámetros

La Tabla 3.19 contiene las *estimaciones de los parámetros asociados a los efectos fijos*. El modelo que estamos ajustando contiene cinco de estos parámetros: la intersección y los cuatro correspondientes a los niveles del factor *tiempo*. De los cuatro parámetros asociados al factor *tiempo*, solo tres son independientes entre sí (recordemos que, dado que los parámetros de efectos fijos se definen como desviaciones de la media total o como desviaciones de unos respecto de otros, su suma vale cero). Por tanto, el modelo que estamos ajustando incluye cuatro parámetros fijos independientes: la intersección y tres de los cuatro parámetros correspondientes a los niveles del factor *tiempo*.

La Tabla 3.19 ofrece las estimaciones de estos cuatro parámetros. El procedimiento fija en cero el correspondiente al último nivel del factor (esta circunstancia se indica en una nota a pie de tabla) y estima los parámetros asociados al resto de niveles por comparación con el que se ha fijado en cero. Al definir los parámetros de esta forma, la constante del modelo (la intersección) toma el valor correspondiente a la media del nivel que se ha fijado en cero.

En nuestro ejemplo, de los cuatro parámetros asociados al factor *tiempo*, el último de ellos (4 = “mes”) se ha fijado en cero y se han estimado los restantes por comparación con él (1 = “hora”, 2 = “día”, 3 = “semana”). El valor estimado para el nivel *hora* (*tiempo* = 1) es la diferencia entre las medias de los niveles *hora* y *mes*: $17 - 7 = 10$ (ver Tabla 3.22). El valor estimado para el nivel *día* (*tiempo* = 2) es la diferencia entre las medias de los niveles *día* y *mes*: $12 - 7 = 5$. Etc. El valor estimado para la *intersección* es la media del nivel que se ha fijado en cero (*mes*). Los resultados de la tabla indican que las medias obtenidas al cabo de una *hora* y de un *día* difieren significativamente de la media obtenida al cabo de un *mes* (*sig.* < 0,0005), y que la diferencia entre las medias obtenidas al cabo de una *semana* y un *mes* no es significativa (*sig.* = 0,254). Los intervalos de confianza permiten llegar a esta misma conclusión, pues únicamente el intervalo correspondiente al nivel *semana* incluye el valor cero.

Tabla 3.19. Estimaciones de los parámetros de efectos fijos

Parámetro	Estimación	Error típ.	gl	t	Sig.	Intervalo de confianza 95%	
						L. inferior	L. superior
Intersección	7,00	1,06	8,29	6,58	,000	4,56	9,44
[tiempo=1]	10,00	,84	15,00	11,86	,000	8,20	11,80
[tiempo=2]	5,00	,84	15,00	5,93	,000	3,20	6,80
[tiempo=3]	1,00	,84	15,00	1,19	,254	-,80	2,80
[tiempo=4]	,00 ^a	,00	.	.	.	.	.

^a. Se ha establecido este parámetro en cero porque es redundante.

La Tabla 3.20 muestra las *estimaciones de los dos parámetros de covarianza* (los dos parámetros asociados a los efectos aleatorios). El modelo de un factor de medidas repetidas incluye dos parámetros de covarianza: la varianza de los *residuos* ($\hat{\sigma}_E^2 = 2,13$) y la varianza de los *sujetos* ($\hat{\sigma}_S^2 = 4,67$). Ambas estimaciones se obtienen a partir de la matriz de varianzas-covarianzas residual **R** (ver Tabla 3.21; en la diagonal principal están

las varianzas muestrales de cada medida repetida; fuera de la diagonal, las covarianzas entre cada par de medidas). Puesto que hemos elegido *simetría compuesta* como estructura de covarianza para la matriz **R**, estamos asumiendo que las cuatro medidas tienen la misma varianza y que la relación entre cualquier par de medidas es la misma (esto es lo que significa *simetría compuesta*). Consecuentemente, los valores de la diagonal principal de **R** son iguales y también son iguales los valores fuera de la diagonal.

En la Tabla 3.20, la varianza de los *residuos* recibe el nombre de *desplazamiento diagonal de SC* porque se obtiene restando al valor de la diagonal principal de **R** (cualquiera de ellos, pues todos son iguales) el valor fuera de esa diagonal (también cualquiera de ellos). La varianza de los *sujetos* recibe el nombre de *Covarianza de SC* porque se corresponde con el valor fuera de la diagonal principal de **R** (cualquiera de ellos), el cual se obtiene promediando las covarianzas entre cada par de medidas repetidas. La sigla *SC* significa *simetría compuesta*, que es la estructura de covarianza que hemos elegido para la matriz residual.

Tabla 3.20. Estimaciones de los parámetros de covarianza

Parámetro	Estim.	Error típico	Wald Z	Sig.	Intervalo de confianza 95%	
					L. inferior	L. superior
Medidas repetidas Desplazam. diagonal de SC	2,13	,78	2,74	,006	1,04	4,36
Covarianza de SC	4,67	3,29	1,42	,157	-1,79	11,12

Tabla 3.21. Matriz de varianzas-covarianzas residual: simetría compuesta

	[tiempo = 1]	[tiempo = 2]	[tiempo = 3]	[tiempo = 4]
[tiempo = 1]	6,80	4,67	4,67	4,67
[tiempo = 2]	4,67	6,80	4,67	4,67
[tiempo = 3]	4,67	4,67	6,80	4,67
[tiempo = 4]	4,67	4,67	4,67	6,80

La varianza de los *residuos* refleja la variabilidad en la calidad del recuerdo no explicada por el paso del tiempo (variabilidad intrasujetos). Y la varianza de los *sujetos* refleja la variabilidad atribuible a las diferencias entre los sujetos (variabilidad intersujetos). El cociente entre la varianza de los sujetos y la varianza total (ver [3.13]) es el **coeficiente de correlación intraclase**⁸:

$$C_{CI} = \hat{\sigma}_S^2 / (\hat{\sigma}_S^2 + \hat{\sigma}_E^2) \quad [3.14]$$

⁸ No confundir este coeficiente con el propuesto en el Capítulo 8 del segundo volumen. La ecuación [3.14] se basa en la variabilidad intersujetos; por tanto, indica el grado de relación existente entre las medidas repetidas (es el C_{CI} que suele utilizarse en *psicometría* para valorar la fiabilidad de las escalas). El coeficiente propuesto en las ecuaciones [8.6] y [8.7] del Capítulo 8 del segundo volumen se basa en la variabilidad intermedidas; por tanto, indica el grado de relación existente entre los sujetos (o proporción de varianza explicada por la diferencia entre las medias de las medidas repetidas).

Este cociente refleja la proporción de la varianza total que es atribuible a la diferencia entre los sujetos; o, de forma equivalente, el grado de parecido o relación existente entre las medidas repetidas. Cuanto mayor es el valor del C_{CI} , más justificado está elegir estructuras de covarianza que no asumen independencia entre las medidas repetidas. En nuestro ejemplo: $C_{CI} = 4,67/(4,67+2,13) = 0,687$. Por tanto, el 68,7% de la varianza de la calidad del recuerdo es atribuible a las diferencias entre los sujetos.

Comparaciones múltiples

La última información solicitada se refiere a las *medias estimadas por el modelo* y a las *comparaciones por pares* entre ellas. La Tabla 3.22 ofrece las medias correspondientes a cada nivel del factor *tiempo*. La tabla incluye, para cada media, su error típico, sus grados de libertad y el intervalo de confianza individual calculado al 95%.

Una vez estimadas las medias, el procedimiento las compara por pares para poder determinar cuáles de ellas difieren entre sí. La Tabla 3.23 incluye, para cada comparación, la diferencia observada entre cada par de medias, el error típico de esa diferencia y el nivel crítico asociado a esa diferencia bajo la hipótesis nula de igualdad de medias (una nota a pie de tabla recuerda que se está aplicando la corrección de Bonferroni para controlar la tasa de error). Los resultados de la tabla indican que, exceptuando la diferencia entre las medias correspondientes a los momentos *semana* y *mes*, todas las diferencias entre medias son significativamente distintas de cero (*sig.* < 0,05 en todos los casos).

Tabla 3.22. Medias estimadas

Tiempo	Media	Error típico	gl	Intervalo de confianza 95%	
				L. inferior	L. superior
Hora	17,00	1,06	8,29	14,56	19,44
Día	12,00	1,06	8,29	9,56	14,44
Semana	8,00	1,06	8,29	5,56	10,44
Mes	7,00	1,06	8,29	4,56	9,44

Tabla 3.23. Comparaciones por pares entre las medias estimadas

(I) Tiempo	(J) Tiempo	Diferencia entre las medias (I-J)	Error típico	gl	Sig. ^a	Intervalo de confianza al 95% para la diferencia ^a	
						L. inferior	L. superior
Hora	Día	5,00	,84	15	,000	2,44	7,56
	Semana	9,00	,84	15	,000	6,44	11,56
	Mes	10,00	,84	15	,000	7,44	12,56
Día	Semana	4,00	,84	15	,002	1,44	6,56
	Mes	5,00	,84	15	,000	2,44	7,56
Semana	Mes	1,00	,84	15	1,000	-1,56	3,56

Basado en las medias marginales estimadas

a. Corrección por comparaciones múltiples: Bonferroni.

Esta conclusión es idéntica a la obtenida al analizar estos mismos datos con el procedimiento **GLM** (ver la Tabla 8.11 del segundo volumen). Sin embargo, los niveles críticos no son idénticos porque han cambiado los errores típicos de las medias estimadas. El procedimiento **GLM** realiza las comparaciones entre cada par de medias utilizando errores típicos que se obtienen a partir de las medias que intervienen en cada comparación. Los errores típicos que calcula el procedimiento **MIXED** dependen de la estructura de covarianza elegida para las medidas repetidas (matriz **R**). Puesto que nosotros hemos elegido *simetría compuesta* (es decir, la misma varianza para todas las medidas repetidas y la misma covarianza entre cada par de medidas repetidas), el procedimiento utiliza el mismo error típico para todas las comparaciones (0,84). Si hubiéramos elegido una matriz **R** *no estructurada* (es decir, sin ningún tipo de estructura predeterminada), los errores típicos habrían sido idénticos a los que ofrece el procedimiento **GLM**.

Análisis de varianza: dos factores con medidas repetidas en ambos

El modelo de dos factores añade al de un factor no solo un factor adicional, sino la interacción entre ambos factores:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + S_i + E_{ijk} \quad [3.15]$$

donde μ es la media de la variable dependiente; α_j es el efecto del primer factor (*A*); β_k es el efecto del segundo factor (*B*); $(\alpha\beta)_{jk}$ es la interacción entre ambos factores; y S_i representa la variabilidad entre las medias de los sujetos. Los E_{ijk} siguen siendo los errores aleatorios.

El modelo asume que S_i y E_{ijk} son variables aleatorias independientes del resto de términos del modelo e independientes entre sí, y distribuidas normalmente con varianzas σ_S^2 y σ_E^2 , respectivamente. Puesto que los nuevos términos β_k y $(\alpha\beta)_{jk}$ tienen, ambos, varianza nula, también en este modelo se verifica

$$\sigma_Y^2 = \sigma_S^2 + \sigma_E^2 \quad [3.16]$$

En el Capítulo 9 del segundo volumen se ofrecen los detalles de este modelo (fuentes de variabilidad) y los efectos que interesa analizar.

Para ajustar un modelo de medidas repetidas a los datos de la Tabla 3.17 (los datos se encuentran en el archivo *Contenido tiempo recuerdo*, el cual puede descargarse de la página web del manual):

- En el cuadro de diálogo previo al principal, trasladar la variable *id* (identificación de caso) a la lista **Sujetos** y las variables *contenido* y *tiempo* a la lista **Repetidas**, seleccionar **Simetría compuesta** en el menú desplegable **Tipo de covarianza para repetidas** y pulsar el botón **Continuar** para acceder al cuadro de diálogo principal.
- Trasladar la variable *recuerdo* (calidad del recuerdo) al cuadro **Variable dependiente** y las variables *tiempo* y *contenido* a la lista **Factores**.

- Pulsar el botón **Fijos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos fijos*, seleccionar las variables *tiempo* y *contenido* y trasladarlas a la lista **Modelo** tras seleccionar **Factorial** en el menú desplegable (el modelo debe incluir los dos efectos principales y el efecto de la interacción). Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Estadísticos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Estadísticos* y marcar las opciones **Estimaciones de los parámetros** y **Contrastes sobre los parámetros de covarianza**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el SPSS ofrece, entre otros, los resultados que muestran las Tablas 3.24 y 3.25 (para revisar los resultados que no se ofrecen en este ejemplo pueden consultarse los ejemplos de los apartados anteriores o el ejemplo del próximo apartado).

Significación de los efectos incluidos en el modelo

La Tabla 3.24 contiene los *contrastes de los efectos fijos*. El modelo propuesto incluye cuatro de estos efectos: (1) la *intersección*, (2) el factor *contenido*, (3) el factor *tiempo* y (4) la interacción *contenido* $\times$ *tiempo*. Cada estadístico *F* de la tabla permite contrastar la hipótesis de que el correspondiente efecto es nulo⁹. Los niveles críticos (*sig.*) indican que los cuatro efectos fijos son significativamente distintos de cero. Por tanto, puede concluirse: (1) que el valor poblacional de la constante o intersección es mayor que cero (*sig.* < 0,0005); (2) que la calidad del recuerdo no es la misma cuando se recuerdan números y cuando se recuerdan letras, es decir, que el tipo de material recordado está relacionado con la calidad del recuerdo (*sig.* < 0,0005); (3) que la calidad del recuerdo no es la misma en los cuatro momentos considerados, es decir, que el paso del

Tabla 3.24. Contraste de los efectos fijos (sumas de cuadrados Tipo III)

Origen	Numerador df	Denominador df	Valor F	Sig.
Intersección	1	5,00	126,78	,000
contenido	1	35	25,77	,000
tiempo	3	35	35,75	,000
contenido * tiempo	3	35	5,17	,005

⁹ Entre estos estadísticos *F* y los que se obtienen con el procedimiento **GLM** (ver Tabla 9.7 en el Capítulo 9 del segundo volumen) existen diferencias de cierta importancia: mientras que el procedimiento **GLM** utiliza medias cuadráticas error calculadas sin asumir que los sujetos son independientes del resto de los efectos fijos presentes en el modelo (se calcula, por tanto, una media cuadrática error para cada efecto fijo; ver Pardo y San Martín, 1998, págs. 356-357), en el procedimiento **MIXED** se asume que los sujetos son independientes del resto de efectos presentes en el modelo; consecuentemente, el procedimiento **MIXED** utiliza una misma media cuadrática error para todos los efectos. No obstante, a pesar de las diferencias existentes tanto en los supuestos que se establecen como en la forma de calcular los estadísticos *F*, ambos procedimientos suelen llevar a la misma conclusión. En el caso de que esto no sea así, la solución del procedimiento **GLM** es preferible a la del procedimiento **MIXED** siempre que la presencia de valores perdidos no constituya un problema importante.

tiempo está relacionado con la calidad del recuerdo ($\text{sig.} < 0,0005$); y (4) que la relación entre el paso del tiempo y la calidad del recuerdo no es la misma al recordar números y al recordar letras ($\text{sig.} = 0,005$).

Estimaciones de los parámetros

La Tabla 3.25 ofrece las *estimaciones de los parámetros de covarianza*. El modelo incluye dos de estos parámetros: (1) la varianza de los *residuos* (*desplazamiento diagonal de SC* = $\hat{\sigma}_E^2 = 1,36$), que refleja la variabilidad de la calidad del recuerdo no explicada por el paso del tiempo y el tipo de material recordado (variabilidad entre las puntuaciones de un mismo sujeto o variabilidad intrasujetos) y (2) la varianza de los *sujetos* (*covarianza de SC* = $\hat{\sigma}_S^2 = 0,95$), que refleja la variabilidad atribuible a las diferencias entre las medias de los sujetos (variabilidad intersujetos). Puesto que se está asumiendo que estas dos varianzas son independientes entre sí y que sumadas agotan la variabilidad total (ver ecuación [3.16]), puede concluirse que las diferencias entre las medias de los sujetos suponen el 41,1 % de la variabilidad total (pues $0,95/(0,95 + 1,36) = 0,411$).

Tabla 3.25. Estimaciones de los parámetros de covarianza

Parámetro		Estimación	Error típico	Wald Z	Sig.
Medidas repetidas	Desplazamiento diag. de SC	1,36	,32	4,18	,000
	Covarianza de SC	,95	,71	1,34	,181

Por supuesto, si se desea obtener comparaciones entre los niveles de los dos factores de medidas repetidas, pueden utilizarse las *comparaciones entre las medias estimadas* estudiadas en el apartado anterior a propósito del modelo de un factor.

También es posible, mediante sintaxis, analizar los efectos simples y realizar las comparaciones necesarias para interpretar el efecto de la interacción (en el siguiente apartado se explica cómo hacer todo esto con la sentencia TEST).

Análisis de varianza: dos factores con medidas repetidas en uno

La formulación matemática de este modelo es idéntica a la del modelo de dos factores con medidas repetidas en ambos estudiado en el apartado anterior (ver ecuación [3.15]). Pero ahora los niveles del factor intersujetos definen varias poblaciones y se asume que la matriz de varianzas-covarianzas residual es la misma en todas ellas. Los detalles de este modelo (fuentes de variabilidad, efectos que interesa analizar, estimaciones del tamaño del efecto, comparaciones múltiples, etc.) pueden consultarse en el Capítulo 9 del segundo volumen.

Para ilustrar cómo ajustar un modelo de dos factores con medidas repetidas en uno de ellos con el procedimiento **MIXED** vamos a utilizar los datos de la Tabla 9.16 del segundo volumen, es decir, vamos a analizar los mismos datos que ya hemos analizado

con el procedimiento **GLM**; de esta forma podremos valorar mejor las diferencias existentes entre ambos procedimientos. Los datos se encuentran en el archivo *Depresión repetidas mixed*, el cual puede descargarse de la página web del manual. Este archivo contiene la misma información que el archivo *Depresión hamilton reducido* utilizado en el segundo volumen, pero organizada tal como requiere el procedimiento **MIXED**:

- En el cuadro de diálogo previo al principal, trasladar la variable *id* (identificación de caso) a la lista **Sujetos** y la variable *momento* a la lista **Repetidas**; seleccionar **Simetría compuesta** en el menú desplegable **Tipo de covarianza para repetidas** y pulsar el botón **Continuar** para acceder al cuadro de diálogo principal.
- Trasladar la variable *hamilton* (puntuaciones escala Hamilton) al cuadro **Variable dependiente** y las variables *tto* y *momento* a la lista **Factores**.
- Pulsar el botón **Fijos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos fijos* y trasladar las variables *tto* y *momento* a la lista **Modelo** tras seleccionar **Factorial** en el menú desplegable. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Estadísticos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Estadísticos* y marcar las opciones **Estimaciones de los parámetros** y **Contrastes sobre los parámetros de covarianza**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Medias marginales estimadas** para acceder al cuadro de diálogo *Modelos lineales mixtos: Medias marginales estimadas* y trasladar la variable *momento* y la interacción *tto* × *momento* a la lista **Mostrar las medias para**. Marcar la opción **Comparar los efectos principales** y, en el menú desplegable **Corrección del intervalo de confianza**, seleccionar **Bonferroni**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Pegar** para generar en el *Editor de sintaxis* las sentencias SPSS correspondientes a las elecciones hechas y modificar la línea “/EMMEANS = TABLES (tto*momento)” añadiendo “COMPARE(tto) ADJ(BONFERRONI)” (la línea debe terminar con un punto). Esta modificación de la sintaxis permite obtener los contrastes de los efectos simples.

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestran las Tablas 3.26 a 3.33.

Significación de los efectos incluidos en el modelo

La Tabla 3.26 muestra los *contrastos de los efectos fijos*. El modelo que estamos ajustando incluye cuatro efectos fijos: (1) la constante del modelo (*intersección*), (2) el factor *tto*, (3) el factor *momento* y (4) la interacción *tto* × *momento*. Cada estadístico *F* de la tabla permite contrastar la hipótesis de que el correspondiente efecto es nulo. Estos estadísticos *F* son idénticos a los que ofrece la opción **Medidas repetidas** del procedi-

miento **GLM** (esfericidad asumida; ver las Tablas 9.22 y 9.23 del segundo volumen). Los resultados obtenidos indican que todos los efectos son distintos de cero (*sig.* < 0,005) con excepción del correspondiente al *tratamiento*¹⁰ (*sig.* = 0,106).

Tabla 3.26. Contraste de los efectos fijos (sumas de cuadrados Tipo III)

Origen	Numerador df	Denominador df	Valor F	Sig.
Intersección	1	38,00	3184,92	,000
tto	1	38,00	2,74	,106
momento	2	76,00	133,42	,000
tto * momento	2	76,00	18,52	,000

Estimaciones de los parámetros

La Tabla 3.27 contiene las medias observadas en las seis casillas resultantes de combinar los dos niveles del factor *tto* y los tres niveles del factor *momento*. Esta tabla de medias servirá para facilitar la interpretación de las estimaciones que ofrece el procedimiento.

Tabla 3.27. Puntuaciones en la escala Hamilton: medias observadas

		Momento		
		Basal	Semana 4	Semana 8
Tratamiento	Estándar	32,65	30,20	27,45
	Combinado	33,85	28,80	22,50

La Tabla 3.28 contiene las *estimaciones de los parámetros asociados a los efectos fijos*. El modelo que estamos ajustando contiene doce de estos parámetros: la intersección, los dos correspondientes a los niveles del factor *tto*, los tres correspondientes a los niveles del factor *momento* y los seis correspondientes a las combinaciones entre los dos niveles del factor *tto* y los tres del factor *momento*. Pero sabemos que no todos estos parámetros son independientes entre sí: hay un solo parámetro independiente asociado al factor *tto*, dos al factor *momento* y dos a la interacción *tto* × *momento*. El resto de parámetros se fijan en cero y únicamente se estiman estos cinco más la intersección (seis estimaciones en total).

- La *intersección* es la puntuación media observada en la escala Hamilton cuando el resto de efectos vale cero. Por tanto, el valor 22,50 es la media observada al cabo de ocho semanas (*momento* = 3) entre los pacientes que han recibido el tratamiento combinado (*tto* = 2).

¹⁰ Este resultado no significa que no existan diferencias entre los tratamientos. El hecho de que el efecto de la interacción *tto* × *momento* sea significativo está indicando que la diferencia entre los tratamientos no es la misma en los tres momentos (en caso necesario, revisar el concepto de interacción en el Capítulo 7 del segundo volumen).

- Las estimaciones correspondientes a cada efecto principal reflejan cómo se desvía la media de cada nivel de la media del nivel fijado en cero, pero solo cuando el otro efecto vale cero. Así, el valor estimado para $tto = 1$ es la diferencia entre las medias de los pacientes que han recibido el tratamiento estándar ($tto = 1$) y los que han recibido el combinado ($tto = 2$) al cabo de ocho semanas ($momento = 3$). En efecto, $27,45 - 22,50 = 4,95$ (ver Tabla 3.27). Tanto el nivel crítico ($sig. < 0,0005$) como el intervalo de confianza (el valor cero no se encuentra entre sus límites) indican que esta diferencia es significativamente distinta de cero.
- Las estimaciones asociadas a los momentos 1 y 2 reflejan las diferencias existentes entre esos niveles y el momento 3. El valor 11,35 indica que, con el tratamiento combinado ($tto = 2$), la media es 11,35 puntos mayor en el momento basal ($momento = 1$) que a las ocho semanas ($momento = 3$). En efecto, $33,85 - 22,50 = 11,35$. Y el valor 6,30 indica que, con el tratamiento combinado ($tto = 2$), la media es 6,30 puntos mayor a las cuatro semanas ($momento = 2$) que a las ocho semanas ($momento = 3$). En efecto, $28,80 - 22,50 = 6,30$. Ambas diferencias son significativamente distintas de cero ($sig. < 0,0005$ en ambos casos).
- Las dos estimaciones correspondientes al efecto de la interacción ($-6,15$ y $-3,55$) reflejan *diferencias entre momentos en cada tratamiento*. El valor $-6,15$ indica que la diferencia entre los momentos 1 y 3 entre quienes han recibido el tratamiento estándar ($32,65 - 27,45 = 5,20$) es 6,15 puntos menor que esa misma diferencia entre quienes han recibido el tratamiento combinado ($33,85 - 22,50 = 11,35$); en efecto, $5,20 - 11,35 = -6,15$. Y el valor $-3,55$ indica que la diferencia entre los momentos 2 y 3 entre quienes han recibido el tratamiento estándar ($30,20 - 27,45 = 2,75$) es 3,55 puntos menor que esa misma diferencia entre quienes han recibido el tratamiento combinado ($28,80 - 22,50 = 6,30$); en efecto, $2,75 - 6,30 = -3,55$. Tanto los niveles críticos ($sig. < 0,05$ en ambos casos) como los intervalos de confianza asociados a estas dos diferencias (ninguno de ellos incluye el valor cero) permiten afirmar que son significativamente distintas de cero.

Tabla 3.28. Estimaciones de los parámetros de efectos fijos

Parámetro	Estimación	Error típico	gl	t	Sig.	Intervalo confianza 95%	
						L. inferior	L. superior
Intersección	22,50	,84	62,95	26,73	,000	20,82	24,18
[tto=1]	4,95	1,19	62,95	4,16	,000	2,57	7,33
[tto=2]	0 ^a	0	.	.	.	.	.
[momento=1]	11,35	,72	76,00	15,82	,000	9,92	12,78
[momento=2]	6,30	,72	76,00	8,78	,000	4,87	7,73
[momento=3]	0 ^a	0	.	.	.	.	.
[momento=1] * [tto=1]	-6,15	1,01	76,00	-6,06	,000	-8,17	-4,13
[momento=2] * [tto=1]	-3,55	1,01	76,00	-3,50	,001	-5,57	-1,53
[momento=3] * [tto=1]	0 ^a	0	.	.	.	.	.
[momento=1] * [tto=2]	0 ^a	0	.	.	.	.	.
[momento=2] * [tto=2]	0 ^a	0	.	.	.	.	.
[momento=3] * [tto=2]	0 ^a	0	.	.	.	.	.

a. Se ha establecido este parámetro en cero porque es redundante.

La Tabla 3.29 muestra las *estimaciones de los parámetros de covarianza*. El modelo que estamos ajustando incluye dos de estos parámetros: la varianza de los *residuos* (*desplazamiento diagonal de SC* = 5,15), que refleja la variabilidad existente dentro de cada sujeto (variabilidad intrasujetos); y la varianza de los *sujetos* (*covarianza de SC* = 9,02), que refleja la variabilidad atribuible a las diferencias entre las medias de los sujetos (variabilidad intersujetos).

Puesto que estamos asumiendo que estas dos varianzas son independientes entre sí y que sumándolas se obtiene la varianza total, el coeficiente de correlación intraclase permite concluir que, una vez controlado el efecto de los factores *tto* y *momento* y de la interacción *tto* $\times$ *momento*, las diferencias entre los sujetos suponen el 64% de la variabilidad total (pues $9,02 / (9,02 + 5,15) = 0,64$).

Tabla 3.29. Estimaciones de los parámetros de covarianza

Parámetro	Estimación	Error típico	Wald Z	Sig.
Medidas repetidas Desplazamiento diag. de SC	5,15	,84	6,16	,000
Covarianza de SC	9,02	2,48	3,64	,000

Comparaciones múltiples

De acuerdo con los resultados de la Tabla 3.26, de los dos efectos principales analizados (*tto* y *momento*) solo es significativo el efecto del factor *momento*. La Tabla 3.30 contiene las medias de cada nivel del factor *momento*, acompañadas de sus errores típicos. Y la Tabla 3.31 ofrece las comparaciones por pares entre esas medias (se han eliminado de la tabla las filas con información redundante, es decir, las comparaciones duplicadas). Para controlar la tasa de error, tanto a los niveles críticos como a los intervalos de confianza se les ha aplicado la corrección de Bonferroni (se recuerda en una nota a pie de tabla).

El resultado de estas comparaciones indica que la media del momento *basal* es significativamente mayor que las medias del resto de los momentos (*sig.* < 0,0005); la media de la semana 4 también es significativamente mayor que la media de la semana 8 (*sig.* < 0,0005). Los intervalos de confianza indican exactamente lo mismo (ninguno de ellos incluye el valor cero). Por tanto, puede concluirse que el nivel medio de depresión (es decir, las puntuaciones medias en la escala Hamilton) va disminuyendo conforme va avanzando el tratamiento. Pero debe tenerse en cuenta que esta conclusión es del todo provisional; el hecho de que el efecto de la interacción sea significativo indica que este resultado podría ser matizado.

Tabla 3.30. Medias estimadas (factor *momento*)

Momento	Media	Error típico	gl
Basal	33,25	,60	62,95
Semana 4	29,50	,60	62,95
Semana 8	24,97	,60	62,95

Tabla 3.31. Comparaciones por pares (factor *momento*)

(I) Momento	(J) Momento	Diferencia entre las medias (I-J)	Error típico	gl	Sig. ^a	Intervalo de confianza al 95% para la diferencia ^a	
						Límite inferior	Límite superior
Basal	Semana 4	3,75	,51	76,00	,000	2,51	4,99
	Semana 8	8,28	,51	76,00	,000	7,03	9,52
Semana 4	Semana 8	4,53	,51	76,00	,000	3,28	5,77

Basado en las medias marginales estimadas

^a. Corrección por comparaciones múltiples: Bonferroni.

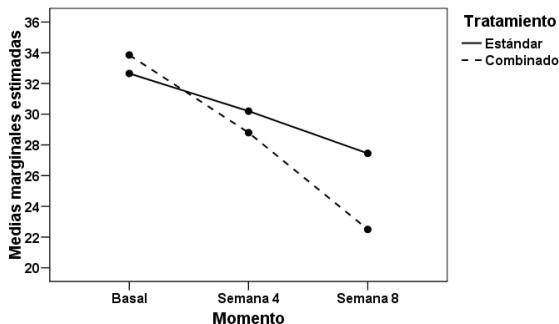
Análisis de los efectos simples

La Tabla 3.32 muestra las medias de las casillas (ver Tabla 3.27), es decir, las medias de cada combinación entre los dos niveles del factor *tto* y los tres niveles del factor *momento* (seis medias en total).

Estas seis medias son las que se utilizan para construir el gráfico de líneas que muestra la Figura 3.3, el cual contiene información útil para entender el significado de los *efectos simples* (distancias verticales entre cada par de puntos) y del *efecto de la interacción* (diferencia entre cada par de distancias verticales).

Tabla 3.32. Medias estimadas (combinaciones *tratamiento* por *momento*)

Momento	Tratamiento	Media	Error típico	gl	Intervalo de confianza 95%	
					Límite inferior	Límite superior
Basal	Estándar	32,65	,84	62,95	30,97	34,33
	Combinado	33,85	,84	62,95	32,17	35,53
Semana 4	Estándar	30,20	,84	62,95	28,52	31,88
	Combinado	28,80	,84	62,95	27,12	30,48
Semana 8	Estándar	27,45	,84	62,95	25,77	29,13
	Combinado	22,50	,84	62,95	20,82	24,18

Figura 3.3. Gráfico de líneas: *tratamiento* por *momento*

La Tabla 3.33 contiene la información relativa a los efectos simples del factor *tto*, es decir, las comparaciones entre los dos niveles del factor *tto* (estándar, combinado) dentro de cada nivel del factor *momento* (basal, semana 4, semana 8). Estas comparaciones aparecen con sus correspondientes pruebas de significación e intervalos de confianza. Una nota a pie de tabla recuerda que se ha aplicado la corrección de Bonferroni tanto a los niveles críticos (*sig.*) como a los intervalos de confianza.

Los resultados obtenidos indican que los tratamientos (el nivel medio de depresión bajo cada tratamiento) difieren significativamente en la semana 8 (*sig.* < 0,0005), pero no en el momento basal ni en la semana 4 (basal: *sig.* = 0,317; semana 4: *sig.* = 0,244). Por tanto, se puede concluir que, en la semana 8, el nivel de depresión es más bajo con el tratamiento combinado que con el estándar; pero no hay evidencia de que esto sea así ni en el momento basal ni en la semana 4.

Tabla 3.33. Comparaciones por pares (efectos simples del factor *tratamiento*)

Momento	(I) Tratam.	(J) Tratam.	Diferencia entre las medias (I-J)	Error típico	gl	Sig. ^a	Intervalo de confianza al 95% ^a	
							L. inferior	L. superior
Basal	Estándar	Combinado	-1,20	1,19	62,95	,317	-3,58	1,18
Semana 4	Estándar	Combinado	1,40	1,19	62,95	,244	-,98	3,78
Semana 8	Estándar	Combinado	4,95	1,19	62,95	,000	2,57	7,33

Basado en las medias marginales estimadas

a. Corrección por comparaciones múltiples: Bonferroni.

Las comparaciones que ofrece la Tabla 3.33 son las que se obtienen como consecuencia de haber modificado la línea de sintaxis “/EMMEANS = TABLES (tto*momento)” añadiendo “COMPARE(tto) ADJ(BONFERRONI)”. Estas mismas comparaciones pueden llevarse a cabo añadiendo a la sintaxis que genera el procedimiento **MIXED** con el botón **Pegar** la siguiente sentencia TEST:

/TEST = ‘Comparaciones entre los dos tratamientos en cada uno de los tres momentos’

```
tto 1 -1 tto*momento 1 0 0 -1 0 0;
tto 1 -1 tto*momento 0 1 0 0 -1 0;
tto 1 -1 tto*momento 0 0 1 0 0 -1.
```

La expresión entre apóstrofes es una etiqueta descriptiva que sirve para recordar lo que estamos intentando hacer. Los códigos 1 y -1 asignados a la variable *tto* indican que se deben comparar los dos niveles de la variable *tto*, es decir, los dos tratamientos. Los códigos asignados a la interacción *tto*momento* indican que esa comparación entre tratamientos debe hacerse dentro de cada nivel del factor *momento*. Para asignar estos códigos debe tenerse en cuenta que las casillas del diseño (las 6 casillas resultantes de combinar los 2 niveles del factor *tto* con los 3 del factor *momento*) se ordenan de la siguiente manera: 1-1, 1-2, 1-3, 2-1, 2-2 y 2-3. Por tanto, los códigos 1 y -1 asignados a la interacción *tto*momento* en la primera línea de la sentencia TEST corresponden a las casillas 1-1 y 2-1, es decir a las dos casillas que contienen las medias del primer fac-

tor en el primer nivel del segundo factor; por tanto, estos códigos están solicitando comparar las medias de los dos tratamientos en el momento basal. Los códigos asignados en la segunda línea ocupan las posiciones correspondientes a las casillas 1-2 y 2-2; por tanto, estos códigos están solicitando comparar las medias de los dos tratamientos en el segundo nivel del segundo factor (semana 4). Finalmente, los códigos asignados en la tercera línea ocupan las posiciones correspondientes a las casillas 1-3 y 2-3; por tanto, estos códigos están solicitando comparar las medias de los dos tratamientos en el tercer nivel del segundo factor (semana 8).

Esta sentencia TEST genera los resultados que muestra la Tabla 3.34, los cuales son idénticos a los ya obtenidos en la Tabla 3.33. Las comparaciones *L1*, *L2* y *L3* se corresponden con las tres líneas de la sentencia: comparaciones entre los tratamientos en el momento basal (*L1*), en la semana 4 (*L2*) y en la semana 8 (*L3*). Los tratamientos únicamente difieren en la semana 8 (*L3*, sig. < 0,0005). Una nota a pie de tabla reproduce la etiqueta descriptiva que hemos incluido en la sintaxis entre apóstrofes.

Tabla 3.34. Efectos simples del factor *tratamiento* en cada nivel del factor *momento* (sentencia TEST)

Contraste ^a	Estimación	Error típico	gl	Valor del contraste	t	Sig.	Intervalo de confianza 95%	
							L. inferior	L. superior
L1	-1,20	1,19	62,95	0	-1,01	,317	-3,58	1,18
L2	1,40	1,19	62,95	0	1,18	,244	-,98	3,78
L3	4,95	1,19	62,95	0	4,16	,000	2,57	7,33

a. Comparaciones entre los dos tratamientos en cada uno de los tres momentos.

Puesto que el factor *tto* únicamente tiene dos niveles, analizar sus efectos simples solo requiere realizar una comparación en cada nivel del factor *momento*: tres comparaciones en total (las tres comparaciones de la Tabla 3.34). Cuando un factor tiene más de dos niveles, además de valorar la significación estadística de cada efecto simple, también puede interesar comparar entre sí las medias involucradas en cada efecto simple. Por ejemplo, el factor *momento* tiene dos efectos simples, uno por cada *tto*; pero cada uno de esos efectos simples incluye tres medias (basal, semana 4 y semana 8). Para precisar el significado de cada uno de estos efectos simples hay que comparar por pares las medias de sus tres niveles (tres comparaciones por tratamiento; seis en total). Estas comparaciones pueden hacerse utilizando dos sentencias TEST, una por cada nivel del factor *tto*. La sintaxis correspondiente es la siguiente:

/TEST = 'Comparaciones por pares entre los tres momentos bajo el tratamiento estándar'

```
momento 1 -1 0 tto*momento 1 -1 0 0 0 0;
momento 1 0 -1 tto*momento 1 0 -1 0 0 0;
momento 0 1 -1 tto*momento 0 1 -1 0 0 0
```

/TEST = 'Comparaciones por pares entre los tres momentos bajo el tratamiento combinado'

```
momento 1 -1 0 tto*momento 0 0 0 1 -1 0;
momento 1 0 -1 tto*momento 0 0 0 1 0 -1;
momento 0 1 -1 tto*momento 0 0 0 0 1 -1.
```

Los códigos asignados al factor *momento* están indicando, en ambas sentencias, que se debe comparar el primer momento con el segundo (primera línea), el primer momento con el tercero (segunda línea) y el segundo momento con el tercero (tercera línea). Los códigos asignados a la interacción *tto*momento* en la primera sentencia están concentrados en el primer tratamiento (*estándar*; casillas 1-1, 1-2 y 1-3); los de la segunda sentencia están concentrados en el segundo tratamiento (*combinado*; casillas 2-1, 2-2 y 2-3).

Las Tablas 3.40 y 3.41 recogen el resultado de estas dos sentencias TEST. Todas las comparaciones entre pares de medias son estadísticamente significativas. Es decir, tanto entre los pacientes que han recibido el tratamiento estándar (Tabla 3.35) como entre los que han recibido el tratamiento combinado (Tabla 3.36) el nivel de depresión es distinto en los tres momentos considerados (las notas a pie de tabla reproducen las etiquetas descriptivas que hemos incluido en la sintaxis entre apóstrofos).

Tabla 3.35. Comparaciones entre *momentos* bajo el tratamiento *estándar* (sentencia TEST)

Contraste ^a	Estimación	Error típico	gl	Valor del contraste	t	Sig.	Intervalo de confianza 95%	
							L. inferior	L. superior
L1	2,45	,72	76,00	0	3,41	,001	1,02	3,88
L2	5,20	,72	76,00	0	7,25	,000	3,77	6,63
L3	2,75	,72	76,00	0	3,83	,000	1,32	4,18

a. Comparaciones por pares entre los tres momentos bajo el tratamiento estándar.

Tabla 3.36. Comparaciones entre *momentos* bajo el tratamiento *combinado* (sentencia TEST)

Contraste ^a	Estimación	Error típico	gl	Valor del contraste	t	Sig.	Intervalo de confianza 95%	
							L. inferior	L. superior
L1	5,05	,72	76,00	0	7,04	,000	3,62	6,48
L2	11,35	,72	76,00	0	15,82	,000	9,92	12,78
L3	6,30	,72	76,00	0	8,78	,000	4,87	7,73

a. Comparaciones por pares entre los tres momentos bajo el tratamiento combinado.

Análisis del efecto de la interacción

La sentencia TEST también sirve para comparar entre sí los efectos simples, es decir, para realizar las comparaciones necesarias para interpretar correctamente una interacción significativa (en caso necesario, revisar, en el Capítulo 7 del segundo volumen, lo relativo al concepto de interacción y a las comparaciones que es necesario llevar a cabo para interpretar una interacción significativa).

Interpretar la interacción requiere comparar entre sí los efectos simples. Para comparar entre sí los efectos simples del factor *tto* (las tres distancias verticales de la Figura 3.3) hay que realizar tres comparaciones: hay que comparar lo que ocurre en el momento basal con lo que ocurre en la semana 4, lo que ocurre en el momento basal con lo que

ocurre en la semana 8 y lo que ocurre en la semana 4 con lo que ocurre en la semana 8). La sentencia TEST para realizar estas tres comparaciones es la siguiente:

/TEST = 'Comparaciones entre los tres efectos simples del factor *tto*'

```
tto*momento 1 -1 0 -1 1 0;
tto*momento 1 0 -1 -1 0 1;
tto*momento 0 1 -1 0 -1 1.
```

Esta sentencia genera los resultados que muestra la Tabla 3.37. Los códigos de la primera línea permiten comparar el primer efecto simple de *tto* con el segundo (la primera distancia vertical de la Figura 3.3 con la segunda). Los códigos de la segunda línea permiten comparar el primer efecto simple de *tto* con el tercero (la primera distancia vertical con la tercera). Los códigos de la tercera línea permiten comparar el segundo efecto simple de *tto* con el tercero (la segunda distancia vertical con la tercera).

Los resultados de la Tabla 3.37 indican que las tres comparaciones solicitadas son significativamente distintas de cero (*sig.* < 0,05 en los tres casos). Por tanto, en lo relativo a las puntuaciones medias en la escala Hamilton, la diferencia entre los dos tratamientos no es la misma en ninguno de los tres momentos considerados. El hecho de que la diferencia entre los tratamientos no sea la misma en la semana 4 y en el momento basal está indicando que, entre esos dos momentos, el nivel de depresión disminuye más con el tratamiento combinado que con el estándar. Y lo mismo está indicando el hecho de que la diferencia entre los tratamientos en la semana 8 no sea la misma que en los dos momentos previos.

Tabla 3.37. Comparaciones entre los efectos simples del factor *tratamiento*

Contraste ^a	Estimación	Error típico	gl	Valor del contraste	t	Sig.	Intervalo de confianza 95%	
							L. inferior	L. superior
L1	-2,60	1,01	76,00	0	-2,56	,012	-4,62	-,58
L2	-6,15	1,01	76,00	0	-6,06	,000	-8,17	-4,13
L3	-3,55	1,01	76,00	0	-3,50	,001	-5,57	-1,53

a. Comparaciones entre los tres efectos simples del factor *tto*.

Análisis de covarianza: dos factores con medidas repetidas en uno

En el apartado anterior hemos visto cómo ajustar un modelo de ANOVA de dos factores con medidas repetidas en uno. En este apartado vamos a ajustar ese mismo modelo, pero incorporando la posibilidad de controlar terceras variables (covariables).

Combinando lo que hemos visto en los apartados anteriores sobre los modelos de medidas repetidas y lo que hemos estudiado en el capítulo anterior sobre el análisis de covarianza se obtiene el modelo de ANCOVA de dos factores con medidas repetidas en uno de ellos:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + S_i + \gamma x_{ijk} + E_{ijk} \quad [3.17]$$

Al igual que en el correspondiente modelo de ANOVA, μ es la media poblacional de la variable dependiente Y ; α_j es el efecto del factor *intersujetos* (factor A); β_k es el efecto del factor *intrasujetos* (factor B); $(\alpha\beta)_{jk}$ es el efecto de la interacción entre ambos factores; S_i representa la variabilidad entre los sujetos; y los E_{ijk} son los errores aleatorios. La diferencia entre este modelo y el correspondiente modelo de ANOVA está en γ , que es el coeficiente de regresión que recoge la relación entre la variable dependiente (Y) y la covariable (X) en cada nivel del factor intersujetos; la letra x minúscula indica que se está trabajando con las puntuaciones centradas.

Se asume que S_i y E_{ijk} son variables aleatorias independientes del resto de los términos del modelo e independientes entre sí, y distribuidas normalmente con varianzas σ_S^2 y σ_E^2 , respectivamente. Los niveles del factor intersujetos definen varias poblaciones cuyas matrices de varianzas-covarianzas residuales se asume que son iguales. También se asume que el efecto de la covariable es independiente del efecto de los factores, es decir, que la pendiente de regresión que relaciona Y con X es la misma en todos los niveles del factor intersujetos. En el caso de que no sea posible asumir esto último, hay que incorporar un nuevo término al modelo:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + S_i + \gamma_{\text{inter}} \bar{x}_{jk} + \gamma_{\text{intra}} x_{ijk} + E_{ijk} \quad [3.18]$$

El coeficiente de regresión *intrasujetos* (γ_{intra}) sigue siendo el coeficiente que recoge la relación entre la covariable y la variable dependiente dentro de cada grupo (dentro de cada nivel del factor intersujetos). El coeficiente de regresión *intersujetos* (γ_{inter}) es el término que permite que las pendientes no sean iguales: recoge la relación entre la variable dependiente y las medias de la covariable en cada nivel del factor intersujetos.

Los modelos [3.17] y [3.18] sirven para analizar los datos provenientes de un diseño con J grupos aleatorios (los J niveles del factor intersujetos A) y K medidas repetidas (los K niveles del factor intrasujetos B). Si no se tienen claros los detalles de este diseño, revisar la Tabla 8.2 del segundo volumen. Al añadir covariables a un diseño de estas características puede procederse de dos maneras distintas: (1) incorporando una sola covariable o (2) incorporando tantas covariables como medidas repetidas.

En el primer caso se mide una única covariable y, consecuentemente, cada sujeto tiene una única puntuación en ella; puesto que cada puntuación X va asociada a K puntuaciones Y , las puntuaciones X deben registrarse antes que las puntuaciones Y . En el segundo caso hay tantas covariables (o mediciones de una misma covariable) como medidas repetidas y, consecuentemente, a cada sujeto le corresponden tantas puntuaciones X como puntuaciones Y ; en este segundo escenario, las puntuaciones X pueden registrarse indistintamente antes o después de su correspondiente puntuación Y . El primer caso puede considerarse un caso especial del segundo en el que las K puntuaciones X de cada sujeto son iguales.

Distinguir entre estos dos escenarios (una covariable, K covariables) es importante por dos razones. La primera de ellas es de índole teórica: en el primer escenario, puesto que la covariable es la misma en todos los niveles del factor intrasujetos, la corrección basada en la covariable únicamente se aplica al efecto intersujetos (es decir, al efecto

del factor *A*); en el segundo escenario, puesto que los valores de la covariable son distintos en todas las condiciones del diseño (es decir, en todas las combinaciones entre los niveles de ambos factores), la corrección basada en la covariable se aplica a todos los efectos del diseño. Por tanto, en el primer escenario (una covariable) los resultados de un ANCOVA serán idénticos a los del correspondiente ANOVA excepto en lo relativo al efecto del factor intersujetos; en el segundo escenario (tantas covariables como medidas repetidas) los resultados de un ANCOVA serán, todos ellos, diferentes de los del correspondiente ANOVA (siempre, claro está, que las covariables estén relacionadas con la variable dependiente).

La segunda razón por la que es importante distinguir entre ambos escenarios es de índole práctica: el procedimiento **GLM** (ver Capítulo 9 del segundo volumen) no permite asociar distintas covariables a las distintas medidas repetidas. Para poder hacer esto es necesario utilizar el procedimiento **MIXED**.

Para ilustrar cómo ajustar un modelo de ANCOVA de dos factores con medidas repetidas en uno de ellos vamos a utilizar el archivo *Depresión repetidas ancova* (puede descargarse de la página web del manual). Este archivo contiene información sobre 379 pacientes sometidos a tratamiento antidepresivo. El objetivo del análisis es valorar el efecto de tres tratamientos (*tto*) y del paso del tiempo (*momento*) sobre las puntuaciones en la escala de depresión de Hamilton (*hamilton*) controlando el efecto de las puntuaciones basales (*cbasal*):

- Utilizar la opción **Seleccionar casos** del menú **Datos** para filtrar las puntuaciones correspondientes a las semanas 2, 4 y 6 (este filtro deja fuera el momento *basal*, lo cual es necesario para poder utilizarlo como covariable).
- Seleccionar la opción **Modelos mixtos > Lineales** del menú **Analizar** para acceder al cuadro de diálogo *Modelos lineales mixtos: Especificar sujetos y medidas repetidas* y trasladar la variable *id* a la lista **Sujetos** y la variable *momento* a la lista **Repetidas**; seleccionar **Simetría compuesta** en el menú desplegable **Tipo de covarianza para repetidas** y pulsar el botón **Continuar** para acceder al cuadro de diálogo principal.
- Trasladar la variable *hamilton* al cuadro **Variable dependiente**, las variables *tto* y *momento* a la lista **Factores** y la variable *cbasal* a la lista **Covariables**.
- Pulsar el botón **Fijos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos fijos*, seleccionar las variables *tto* y *momento* y trasladarlas a la lista **Modelo** tras seleccionar **Factorial** en el menú desplegable; trasladar a continuación la variable *cbasal* (el modelo debe incluir los efectos principales *tto* y *momento*, la interacción entre *tto* y *momento* y la covariable *cbasal*). Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestra la Tabla 3.38. La tabla contiene los contrastes de los cuatro efectos fijos que incluye el modelo propuesto: (1) el efecto del factor intersujetos *tto*, (2) el efecto del factor intrasujetos *momento*, (3) el efecto de la interacción entre *tto* y *momento*, y (4) el efecto de la covariable *cbasal*.

La covariable *cbasal* está relacionada con la variable dependiente ($\text{sig.} < 0,0005$). Por tanto, parece que tiene sentido haberla incluido en el análisis para controlar su efecto. Recordemos que este control afecta únicamente al efecto de los tratamientos (factor intersujetos): puesto que el valor de la covariable es el mismo en todos los niveles intrasujetos, el factor *momento* y la interacción *tto* $\times$ *momento* (los dos efectos intrasujetos que incluye el modelo) no se ven alterados por la presencia de la covariable.

Los resultados de la Tabla 3.38 indican que, una vez controlado el efecto de las puntuaciones basales: (1) el nivel de depresión (es decir, las puntuaciones medias en la escala Hamilton) no es el mismo en los tres *tratamientos* (en el mismo análisis dejando fuera la covariable *cbasal*, es decir, en el correspondiente ANOVA, el efecto de los tratamientos también es significativo: $F = 5,42$; $\text{sig.} = 0,005$); (2) el nivel de depresión no es el mismo en los tres *momentos*; y (3) las diferencias entre los tratamientos no son iguales en los tres momentos.

Tabla 3.38. Contraste de los efectos fijos (ANCOVA)

Origen	Numerador df	Denominador df	Valor F	Sig.
Intersección	1	375,00	7785,32	,000
tto	2	375,00	30,77	,000
momento	2	752,00	940,18	,000
tto * momento	4	752,00	41,08	,000
cbasal	1	375,00	674,02	,000

Estructura de la matriz de varianzas-covarianzas residual

En los modelos de ANOVA que permite ajustar la opción **Modelo lineal general > Medidas repetidas** (procedimiento **GLM**; ver Capítulos 8 y 9 del segundo volumen) se asume que los errores son independientes entre sí y que se distribuyen $N(0, \mathbf{R})$, donde $\mathbf{R}$ es una matriz de varianzas-covarianzas desconocida que se asume que es *esférica* (ver Capítulo 8 del segundo volumen si se necesita aclarar el significado de este supuesto). En los ejemplos de los apartados anteriores de este mismo capítulo, al elegir *simetría compuesta* como estructura de covarianza para la matriz $\mathbf{R}$ hemos hecho algo parecido: hemos supuesto que las varianzas poblacionales son iguales e iguales también las covarianzas entre cada par de medidas repetidas.

No obstante, cuando se trabaja con medidas repetidas parece razonable asumir, no solo que las diferentes medidas no son independientes entre sí, sino que las medidas que están más cercanas en el tiempo podrían correlacionar más que las que están más alejadas. Esta circunstancia puede incorporarse al análisis eligiendo una estructura de covarianza que represente lo mejor posible las relaciones subyacentes. El menú desplegable **Tipo de covarianza para repetidas**, ubicado en el cuadro de diálogo previo al principal, permite elegir entre diferentes estructuras de covarianza. La ayuda del procedimiento ofrece una descripción detallada de todas ellas.

Si no se indica otra cosa, el procedimiento **MIXED** utiliza, para las medidas repetidas, una matriz de varianzas-covarianzas (matriz **R**; ver Apéndice 3) de tipo *diagonal*. La Tabla 3.39 muestra la matriz diagonal correspondiente al ejemplo utilizado en el apartado *Análisis de varianza: un factor con medidas repetidas* (sobre la relación entre el paso del tiempo y la calidad del recuerdo). Una matriz diagonal contiene las varianzas muestrales en la diagonal principal y ceros fuera de la diagonal. En este tipo de estructura de covarianza no se está asumiendo que las varianzas poblacionales de las medidas repetidas (los valores de la diagonal principal) son iguales, pero sí que la relación entre cada par de medidas repetidas (los valores fuera de la diagonal principal) es nula.

Con un factor de medidas repetidas no es razonable asumir que los niveles del factor (las medidas repetidas) son independientes entre sí. Esta es la razón por la que en los ejemplos de los apartados anteriores hemos cambiado la estructura de covarianza que el procedimiento utiliza por defecto (*diagonal*) por una opción que asume que las medidas repetidas están relacionadas (*simetría compuesta*). Al elegir simetría compuesta (ver Tabla 3.21) se está asumiendo que las varianzas de las medidas repetidas son iguales entre sí (valores iguales en la diagonal principal) e iguales entre sí también las covarianzas entre cada par de medidas (valores iguales fuera de la diagonal principal, pero no necesariamente ceros). Esta estructura de covarianza es muy útil porque es la más parsimoniosa de todas (incluye el menor número posible de parámetros), pero no siempre es la mejor elección cuando se trabaja con medidas repetidas. El procedimiento **MIXED** permite elegir otras estructuras de covarianza.

Tabla 3.39. Matriz de varianzas-covarianzas residual: *diagonal*

	[tiempo = 1]	[tiempo = 2]	[tiempo = 3]	[tiempo = 4]
[tiempo = 1]	4,80	0	0	0
[tiempo = 2]	0	8,40	0	0
[tiempo = 3]	0	0	6,80	0
[tiempo = 4]	0	0	0	7,20

La Tabla 3.40 muestra la matriz *sin estructura* correspondiente a nuestro ejemplo sobre la relación entre el paso del tiempo y la calidad del recuerdo (ver el apartado *Análisis de varianza: un factor con medidas repetidas*). Al igual que el resto de matrices de varianzas-covarianzas, es una matriz simétrica: la relación entre los momentos 1 y 2 es la misma que entre los momentos 2 y 1; etc. Una matriz sin estructura admite cualquier pauta de asociación entre las medidas repetidas, pero esta versatilidad se consigue a costa de utilizar el mayor número posible de parámetros. Por tanto, aunque con una matriz sin estructura se consigue siempre el mejor ajuste posible a los datos (por el mayor número de parámetros de covarianza que incluye), el modelo que se genera es el menos parsimonioso de todos. Casi siempre hay otro tipo de estructura más simple y, casi siempre, teóricamente más justificable.

Una forma razonable de proceder para elegir la estructura de covarianza idónea consiste en comenzar obteniendo una matriz *sin estructura* y estudiar detenidamente las

pautas de variación presentes en ella para averiguar si se ajusta o parece a alguna estructura de covarianza más simple. La estructura de covarianza idónea debe guardar un equilibrio entre posibilitar el mejor ajuste posible a los datos (criterio de ajuste) y, al mismo tiempo, ser lo más simple posible para no tener que estimar más parámetros de los estrictamente necesarios (criterio de parsimonia).

En el ejemplo de la Tabla 3.40 (matriz *sin estructura*) las varianzas no son muy diferentes entre sí y tampoco parece que su valor vaya aumentando o disminuyendo de forma evidente entre un momento y otro; por tanto, no es descabellado asumir que las correspondientes varianzas poblacionales son iguales. Tampoco las covarianzas parece que aumenten o disminuyan al aumentar o disminuir la distancia temporal entre las medidas. En principio, por tanto, una estructura de *simetría compuesta* parece una elección razonable.

Tabla 3.40. Matriz de varianzas-covarianzas residual: *sin estructura*

	[tiempo = 1]	[tiempo = 2]	[tiempo = 3]	[tiempo = 4]
[tiempo = 1]	4,80	5,00	4,20	4,40
[tiempo = 2]	5,00	8,40	6,00	3,80
[tiempo = 3]	4,20	6,00	6,80	4,60
[tiempo = 4]	4,40	3,80	4,60	7,20

Una estructura de covarianza frecuentemente utilizada cuando se trabaja con medidas repetidas es la *autorregresiva*. En una estructura de este tipo, las medidas más próximas correlacionan entre sí más que las más alejadas. En la estructura autorregresiva de primer orden (AR1), la covarianza entre medidas separadas k posiciones es $\sigma_Y^2 \rho^k$, donde σ_Y^2 se refiere a la varianza total y ρ al coeficiente de correlación intraclase (ver Tabla 3.41).

El procedimiento **MIXED** ofrece las estimaciones de ambos parámetros cuando se elige AR1 como estructura de covarianza para las medidas repetidas; en nuestro ejemplo tenemos $\hat{\sigma}_Y^2 = 6,522$ y $\hat{\rho} = 0,718$. En la Tabla 3.41 se puede apreciar que las cuatro varianzas (los valores de la diagonal principal) son iguales entre sí y que el valor de las covarianzas (los valores fuera de la diagonal) va disminuyendo conforme las medidas van estando más separadas. Entre las medidas adyacentes (las medidas separadas una sola posición) la covarianza vale $6,522(0,718)^1 = 4,68$; entre las medidas separadas dos posiciones vale $6,522(0,718)^2 = 3,36$; y entre las medidas separadas tres posiciones, vale $6,522(0,718)^3 = 2,42$.

Tabla 3.41. Matriz de varianzas-covarianzas residual: AR1 (autorregresiva de primer orden)

	[tiempo = 1]	[tiempo = 2]	[tiempo = 3]	[tiempo = 4]
[tiempo = 1]	6,52	4,68	3,36	2,42
[tiempo = 2]	4,68	6,52	4,68	3,36
[tiempo = 3]	3,36	4,68	6,52	4,68
[tiempo = 4]	2,42	3,36	4,68	6,52

La estructura AR1 es un caso particular de una estructura más general llamada *Toeplitz*¹¹. En esta estructura, el valor de las covarianzas depende de la distancia entre las medidas repetidas, pero no se obtienen a partir de un único coeficiente de correlación intraclass, como en AR1, sino que se estima un coeficiente de correlación para cada distancia (ver Tabla 3.42).

Tabla 3.42. Matriz de varianzas-covarianzas residual: *Toeplitz*

	[tiempo = 1]	[tiempo = 2]	[tiempo = 3]	[tiempo = 4]
[tiempo = 1]	6,81	4,99	4,02	5,08
[tiempo = 2]	4,99	6,81	4,99	4,02
[tiempo = 3]	4,02	4,99	6,81	4,99
[tiempo = 4]	5,08	4,02	4,99	6,81

Todos estos ejemplos permiten constatar que existen diversas maneras de configurar la estructura de covarianza de la matriz residual (matriz **R**). Aquí hemos prestado atención a cinco de ellas: diagonal, simetría compuesta, sin estructura, autorregresiva de primer orden y Toeplitz. Puesto que las diferencias entre ellas son evidentes, ¿cuál elegir?

La estructura de covarianza elegida afecta tanto al grado de ajuste del modelo como al número de parámetros que es necesario estimar. Y ya hemos señalado que el criterio que debe guiar la elección de la estructura idónea es el equilibrio entre el máximo ajuste posible y el mínimo número de parámetros. La Tabla 3.43 ofrece los estadísticos de ajuste obtenidos y el número de parámetros que es necesario estimar con cada una de las cinco estructuras de covarianza mencionadas. Exceptuando la desviación ($-2LL$), el resto de los estadísticos penalizan el ajuste con alguna función del número de parámetros que es necesario estimar. Aunque un modelo se ajusta tanto mejor cuanto más parámetros tiene, esto no quiere decir que un modelo con más parámetros sea mejor: lo ideal es encontrar el modelo capaz de conseguir un buen ajuste con el menor número de parámetros.

El estadístico $-2LL$ indica que el modelo que mejor ajuste ofrece es el que no utiliza ninguna estructura de covarianza: puesto que el valor de $-2LL$ no se ve afectado por el número de parámetros estimados, la matriz *sin estructura* siempre es la que mejor ajuste ofrece. Del resto de modelos, el que mejor ajuste ofrece es el que utiliza una estructura *Toeplitz*. No obstante, al revisar los estadísticos de bondad de ajuste que penalizan por el número de parámetros estimados, los modelos que utilizan *simetría compuesta* y AR1 son los que ofrecen, sistemáticamente, mejor resultado. Dado que ambas estructuras son, además, las más parsimoniosas, la elección idónea podría recaer sobre cualquiera de esas dos.

Conviene saber que, aunque las estimaciones de los efectos fijos apenas se ven alteradas por el tipo de estructura de covarianza elegida para las medidas repetidas, no ocu-

¹¹ Tanto esta estructura más general como la más particular AR1 solo tiene sentido utilizarlas si los niveles del factor (las medidas repetidas) están igualmente espaciados. En nuestro ejemplo, las medidas no están igualmente espaciadas (hora, día, semana, mes); se aplican estas estructuras únicamente para ilustrar su uso.

re lo mismo con sus errores típicos. Esto implica que los estadísticos y niveles críticos utilizados para tomar decisiones sobre los efectos fijos pueden cambiar dependiendo de la muestra concreta utilizada. Y aunque estos cambios suelen ser poco importantes, es muy recomendable vigilarlos (particularmente cuando se obtienen resultados no previstos).

Tabla 3.43. Número de parámetros y estadísticos de ajuste con distintas estructuras de covarianza

<i>Estructura de covarianza</i>	<i>Nº de parámetros de covarianza estimados</i>	<i>-2LL</i>	<i>AIC</i>	<i>AICC</i>	<i>CAIC</i>	<i>BIC</i>
Sin estructura	10	86,11	106,11	130,55	126,06	116,06
Toeplitz	4	88,27	96,27	98,94	104,26	100,26
Diagonal	4	101,86	109,86	112,53	117,85	113,85
Simetría compuesta	2	90,46	94,46	95,17	98,46	96,46
AR1	2	90,55	94,55	95,26	98,54	96,54

Apéndice 3

Elementos de un modelo lineal mixto

Un modelo lineal que únicamente incluye efectos fijos (al margen de los errores) adopta, en notación matricial, la siguiente forma:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{E} \quad [3.19]$$

donde $\mathbf{Y}$ es un vector columna de orden $n \times 1$ que contiene las puntuaciones de la variable dependiente; $\boldsymbol{\alpha}$ es un vector columna de orden $(p+1) \times 1$ que contiene los *parámetros de efectos fijos*: el término constante μ y un término más ($\alpha_1, \alpha_2, \dots, \alpha_j, \dots, \alpha_p$) por cada una de las p variables independientes; $\mathbf{X}$ es la *matriz del diseño* para los efectos fijos: una matriz de orden $n \times (p+1)$ que contiene las puntuaciones de las p variables independientes más un vector de unos en la primera columna para recoger el efecto del término constante; y $\mathbf{E}$ es un vector columna de orden $n \times 1$ que contiene los errores del modelo (es decir, la parte de $\mathbf{Y}$ que no está explicada por $\mathbf{X}\boldsymbol{\alpha}$). En el modelo propuesto en [3.19] se asume que los errores son independientes entre sí y que se distribuyen normalmente con media 0 y varianza σ_E^2 .

Puesto que una distribución normal queda completamente especificada fijando el valor de su media y el de su varianza, para estimar los parámetros de un modelo lineal como el propuesto en [3.19] mediante métodos que asumen normalidad (como, por ejemplo, los métodos de máxima verosimilitud), basta con asumir que los errores (único término aleatorio del modelo) se distribuyen normalmente con media 0 y varianza σ_E^2 . Puesto que los errores son la única fuente aleatoria

que actúa sobre la variable dependiente, al asumir que se distribuyen normalmente con varianza σ_E^2 , la distribución de $\mathbf{Y}$ queda completamente especificada: es normal con $\text{Var}(\mathbf{Y}) = \sigma_E^2 \mathbf{I}$.

Cuando un modelo lineal contiene una mezcla de términos de efectos fijos y de efectos aleatorios, se tiene un modelo lineal mixto:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} + \mathbf{E} \quad [3.20]$$

$\mathbf{Z}$ es la *matriz del diseño*, de orden $n \times q$, para los efectos aleatorios (se define igual que $\mathbf{X}$ con la diferencia de que $\mathbf{Z}$ no incluye el vector inicial de unos) y $\boldsymbol{\beta}$ es un vector de orden $q \times 1$ que contiene los *parámetros de efectos aleatorios*. En el modelo propuesto en [3.20] se está asumiendo que $\boldsymbol{\beta}$ y $\mathbf{E}$ son independientes entre sí y que se distribuyen normalmente con media 0 y varianzas $\mathbf{G}$ y $\mathbf{R}$, respectivamente. Cuando $\mathbf{Z} = \mathbf{0}$ y $\mathbf{R} = \sigma_E^2 \mathbf{I}$, el modelo [3.20] se reduce al modelo lineal de efectos fijos propuesto en [3.19].

Un modelo mixto contiene dos partes: la referida a los efectos fijos y la referida a los efectos aleatorios. Los parámetros asociados a los efectos fijos ($\boldsymbol{\alpha}$) se consideran constantes fijas; los asociados a los efectos aleatorios ($\boldsymbol{\beta}$) se consideran variables aleatorias. Por tanto, las fuentes aleatorias que actúan sobre la variable dependiente de un modelo lineal mixto son dos: la que se deriva de los parámetros de efectos aleatorios ($\boldsymbol{\beta}$) y la que se deriva del término error ($\mathbf{E}$). Consecuentemente, la varianza de $\mathbf{Y}$ dependerá tanto de $\mathbf{G}$ (la varianza de los efectos aleatorios) como de $\mathbf{R}$ (la varianza de los errores). En concreto: $\text{Var}(\mathbf{Y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$. En efecto,

$$\text{Var}(\mathbf{Y}) = \text{Var}(\mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} + \mathbf{E})$$

y dado que se está asumiendo que todos los términos del modelo son independientes entre sí, se verifica

$$\text{Var}(\mathbf{Y}) = \text{Var}(\mathbf{X}\boldsymbol{\alpha}) + \text{Var}(\mathbf{Z}\boldsymbol{\beta}) + \text{Var}(\mathbf{E})$$

Ahora bien, $\boldsymbol{\alpha}$ contiene los parámetros de efectos fijos; por tanto, $\text{Var}(\boldsymbol{\alpha}) = \mathbf{0}$. Y dado que $\mathbf{X}$ y $\mathbf{Z}$ son matrices de constantes,

$$\text{Var}(\mathbf{Y}) = \mathbf{Z}[\text{Var}(\boldsymbol{\beta})]\mathbf{Z}' + \text{Var}(\mathbf{E}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} \quad [3.21]$$

Para que la distribución de la variable dependiente $\mathbf{Y}$ quede completamente especificada es necesario asumir una determinada estructura tanto para $\mathbf{G}$ como para $\mathbf{R}$. Una de las estructuras de covarianza más simples y utilizadas es la conocida como *componentes de la varianza*. En esta estructura, la matriz $\mathbf{G}$ es una matriz diagonal que contiene los componentes de la varianza de los términos aleatorios del modelo y la matriz $\mathbf{R}$ es proporcional a una matriz identidad: $\mathbf{R} = \sigma_E^2 \mathbf{I}$. Pero ésta no es la única estructura de covarianza disponible. El procedimiento **Mixed** ofrece la posibilidad de definir y utilizar diferentes estructuras de covarianza tanto para $\mathbf{G}$ como para $\mathbf{R}$ (ver, en este mismo capítulo, el apartado *Estructura de la matriz de varianzas-covarianzas residual*).

Métodos de estimación en los modelos lineales mixtos

Dadas las peculiaridades de los modelos mixtos, con el método de *mínimos cuadrados*, que es el método de estimación que se utiliza en los modelos de análisis de varianza y de regresión lineal, no siempre es posible estimar los parámetros de un modelo mixto y, cuando es posible hacerlo, no siempre se obtienen estimaciones óptimas. Esto es particularmente cierto cuando los parámetros se definen como variables aleatorias (no como constantes fijas) y cuando los errores

no son independientes entre sí. Con los modelos mixtos y con los modelos lineales generalizados que estudiaremos en los próximos capítulos es preferible realizar estimaciones mediante el método de máxima verosimilitud.

El procedimiento **MIXED** del SPSS incluye dos versiones de este método: **máxima verosimilitud** (MV) y **máxima verosimilitud restringida** (MVR) (puede encontrarse una buena descripción de estos métodos en Brown y Prescott, 1999, págs. 44-55; o en Verbeke y Molenberghs, 2000, págs. 41-47). Para ajustar e interpretar correctamente un modelo mixto con un programa informático no es necesario conocer cómo funcionan los métodos de estimación; sin embargo, nos parece que no está de más mencionar brevemente en qué consisten.

En el procedimiento **MIXED**, el método MV asume que los datos se ajustan a una distribución normal (la estimación por máxima verosimilitud requiere trabajar con una distribución conocida). Las estimaciones de un grupo de parámetros se obtienen maximizando la función de verosimilitud respecto de ese grupo de parámetros, el cual está formado por todos los efectos fijos, incluida la constante del modelo, y todos los efectos aleatorios. Las estimaciones de máxima verosimilitud son los valores en los que el logaritmo de la función de verosimilitud alcanza su máximo local. El procedimiento calcula las estimaciones de máxima verosimilitud aplicando un algoritmo iterativo que combina el método de Newton-Raphson y el método de *tanteo* (*scoring*) de Fisher. Este algoritmo funciona realizando cálculos de forma repetida hasta alcanzar determinados criterios preestablecidos. En la primera iteración se utiliza el método de Fisher; en las demás iteraciones se utiliza el de Newton-Raphson (cuando existen problemas de convergencia, éstos suelen resolverse haciendo que el método de Fisher actúe en las dos, tres, ..., primeras iteraciones). Para conocer los detalles de estos algoritmos puede consultarse Green, 1984; Jennrich y Sampson, 1976; o Searle, Casella y McCulloch, 1992, pág. 295.

El método MVR es, en esencia, idéntico al de máxima verosimilitud. La única diferencia está en que en la versión restringida se tienen en cuenta los grados de libertad utilizados para estimar los parámetros correspondientes a los efectos fijos. En lugar de usar el vector original de datos, el método MVR se basa en combinaciones lineales de los datos elegidas de tal forma que sean invariantes para los parámetros de efectos fijos incluidos en el modelo. De este modo, la maximización se lleva a cabo sobre un vector restringido. Para conocer en detalle cómo funciona este método puede consultarse Corbeil y Searle (1976), McCulloch y Searle (2001, págs. 21, 176-178) o Searle, Casella y McCulloch (1992).

4

Modelos lineales multinivel

Los modelos mixtos estudiados en el capítulo anterior engloban un tipo particular de modelos lineales llamados de *coeficientes aleatorios* (Longford, 1993), *jerárquicos* (Raudenbush y Bryk, 2002) o *multinivel* (Goldstein, 2003). Son modelos apropiados para analizar datos cuando los casos están anidados en unidades de información más amplias (grupos) y se efectúan mediciones tanto en el nivel más bajo (los casos) como en los niveles más altos (los grupos).

Las estructuras jerárquicas o multinivel se dan en muchos contextos: los pacientes están agrupados en centros hospitalarios; los alumnos están agrupados en aulas y éstas en colegios; los individuos están agrupados en familias, éstas en barrios y éstos en ciudades; etc.

Desde el punto de vista del análisis de datos, el hecho relevante de este tipo de estructuras es que los pacientes del mismo centro hospitalario, o los alumnos del mismo colegio, o los individuos de la misma familia cabe esperar que sean más *parecidos entre sí* que los pacientes de distintos centros hospitalarios, o los alumnos de diferentes colegios, o los individuos de diferentes familias. Esto significa que los sujetos que pertenecen al mismo subgrupo no son, muy probablemente, independientes entre sí; y esto constituye un serio incumplimiento de un supuesto básico del modelo lineal general: la independencia entre observaciones. Los modelos lineales mixtos permiten abordar este tipo de estructuras multinivel prestando atención a la covarianza existente en los datos.

En nuestro ejemplo sobre pacientes afectados de trastorno depresivo (archivo *Depresión*) existen variables propias de los pacientes (nivel 1) y variables propias de los centros (nivel 2). Las puntuaciones en la escala de Hamilton o el sexo son variables medidas en el nivel 1; el tipo de centro (público, privado) o la edad media de los pacien-

tes de cada centro son variables medidas en el nivel 2 (la edad de cada paciente es una variable del nivel 1, pero la edad media de cada centro es una variable del nivel 2). En este capítulo se describen algunos de los modelos multinivel más utilizados (ver Bickel, 2007; Goldstein, 2003; Heck y Thomas, 2000; Hox, 2010; Luke, 2004; Raudenbush y Bryk, 2002).

Qué es un modelo multinivel

Para entender en qué consiste un modelo multinivel conviene comenzar estudiando la relación entre dos variables, por ejemplo, las puntuaciones basales en la escala Hamilton (variable independiente X) y el grado de recuperación al cabo de seis semanas (variable dependiente Y) en un centro hospitalario cualquiera. Ambas son variables del nivel 1, es decir, tanto las puntuaciones basales como la recuperación se refieren a cada paciente individualmente considerado. La ecuación de regresión lineal que expresa la relación entre estas dos variables adopta la forma (ver Capítulo 1):

$$Y_i = \beta_0 + \beta_1 X_i + E_i \quad [4.1]$$

El coeficiente β_0 (la *constante* o *intersección*) es la recuperación pronosticada para un paciente cuya puntuación basal vale cero. El coeficiente β_1 (la pendiente de la recta de regresión) es el cambio pronosticado en la variable dependiente Y (la recuperación) por cada unidad que aumenta la variable independiente X (las puntuaciones basales). El término E_i representa el error asociado a cada pronóstico individual (la diferencia entre la recuperación real y la pronosticada por el modelo de regresión). Se asume que estos errores se distribuyen normalmente con varianza σ_E^2 . La Figura 4.1 (izquierda) muestra la nube de puntos y la ecuación de regresión de Y sobre X en un hipotético centro hospitalario. La pendiente de la recta (positiva) indica que las puntuaciones basales más altas (más bajas) tienden a ir acompañadas de mayor (menor) recuperación.

Para que el coeficiente β_0 tenga un significado claro es habitual re-escalar los valores de la variable independiente. Así, por ejemplo, restando a cada puntuación basal su media (es decir, utilizando las puntuaciones diferenciales o centradas en lugar de las directas), el coeficiente β_0 se convierte en la *media* de la variable dependiente, que es justamente el pronóstico correspondiente a la puntuación basal media:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \quad (\text{con } x_i = X_i - \bar{X}) \quad [4.2]$$

La Figura 4.1 (derecha) muestra la nube de puntos y la ecuación de regresión con las puntuaciones basales *centradas* (al centrar X cambia el valor de β_0 , pero no la forma de la nube de puntos ni la pendiente β_1).

Consideremos ahora dos centros hospitalarios distintos. La Figura 4.2 ilustra cómo se comporta la relación entre la recuperación, Y , y las puntuaciones basales centradas, x , en dos centros hipotéticos (círculos y triángulos). Los dos centros (las dos rectas de regresión) representados en el gráfico de la izquierda únicamente difieren en la recupe-

Figura 4.1. Relación entre la *recuperación* y las *puntuaciones basales* (izquierda) y las *puntuaciones basales centradas* (derecha) en un hipotético centro hospitalario

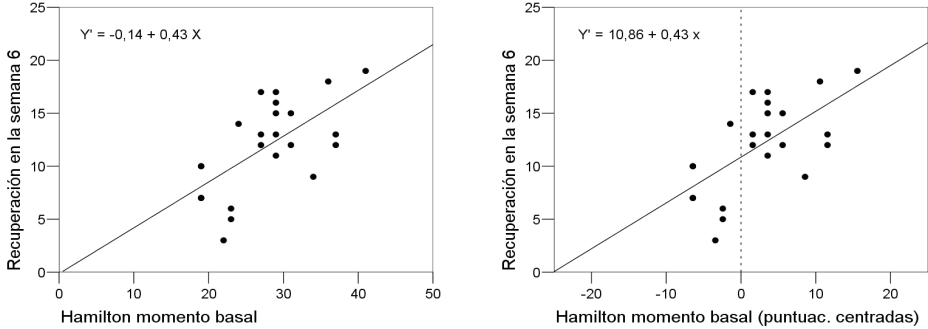
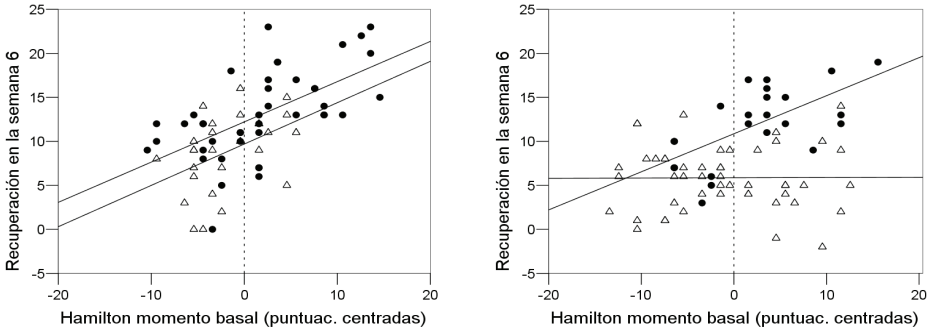


Figura 4.2. Relación entre la *recuperación en la semana 6* y las *puntuaciones basales centradas*. Cada recta de regresión se refiere a un centro hospitalario distinto



ración media (β_0): la media del centro representado con círculos es mayor que la del representado con triángulos; sin embargo, sus pendientes (β_1) son prácticamente idénticas. Por el contrario, los dos centros representados en el gráfico de la derecha difieren tanto en sus medias como en sus pendientes: el centro representado con círculos tiene mayor media y mayor pendiente que el representado con triángulos. Para reflejar estas diferencias entre los dos centros es necesario recurrir a dos ecuaciones de regresión distintas, una para cada centro:

$$\begin{aligned} Y_{i1} &= \beta_{01} + \beta_{11}x_{i1} + E_{i1} \\ Y_{i2} &= \beta_{02} + \beta_{12}x_{i2} + E_{i2} \end{aligned} \quad (\text{con } x_i = X_i - \bar{X}) \quad [4.3]$$

(el subíndice j se refiere a los centros: $j = 1, 2$). La primera ecuación (Y_{i1}) recoge la relación entre la recuperación y las puntuaciones basales en el centro 1; la segunda ecuación (Y_{i2}) recoge esa misma relación en el centro 2. Puesto que la variable x está centrada, el coeficiente β_{01} representa la recuperación media de los pacientes del cen-

tro 1; el coeficiente β_{02} , la de los pacientes del centro 2. El coeficiente β_{11} es la pendiente del centro 1; el coeficiente β_{12} , la del centro 2; ambas pendientes representan el cambio pronosticado en la recuperación de los pacientes por cada unidad que aumentan las puntuaciones basales.

Si se tienen J centros en lugar de dos, no es necesario recurrir a J ecuaciones de regresión; es más práctico utilizar una sola ecuación para todos los centros:

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + E_{ij} \quad [4.4]$$

(por simplicidad se asume que los errores E_{ij} se distribuyen normalmente y con igual varianza en todos los centros). Ahora, tanto la intersección como la pendiente aparecen con el subíndice j , lo cual significa que el modelo permite a cada centro tener su propia intersección y su propia pendiente¹. Y justamente esta variabilidad en el segundo nivel es lo que caracteriza a un modelo multinivel: la ecuación propuesta en [4.4] permite modelar cómo se relacionan las unidades del primer nivel (los pacientes) en cada uno de los subgrupos definidos por la variable del segundo nivel (los centros).

Lo que interesa destacar en este momento es que los parámetros β_{0j} y β_{1j} ya no se interpretan como constantes fijas, como en el modelo de regresión clásico, sino como *variables* cuyos valores pueden cambiar de un centro a otro:

$$\begin{aligned} \beta_{0j} &= \gamma_0 + U_{0j} \\ \beta_{1j} &= \gamma_1 + U_{1j} \end{aligned} \quad [4.5]$$

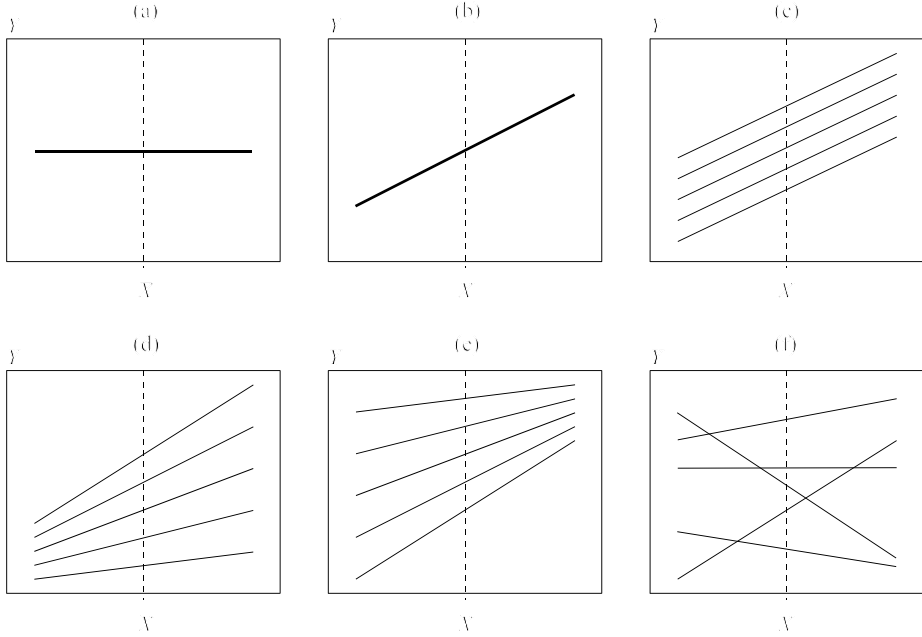
Es decir, el coeficiente β_{0j} está formado por (1) una parte fija o sistemática, γ_0 , que representa la recuperación media en la población de centros y (2) una parte aleatoria, U_{0j} , que representa la variabilidad de las medias de los distintos centros en torno a la media global γ_0 . Del mismo modo, el término β_{1j} está formado por (1) una parte fija o sistemática, γ_1 , que es la pendiente media que relaciona la recuperación y las puntuaciones basales en la población de centros y (2) una parte aleatoria, U_{1j} , que representa la variabilidad de las pendientes de los distintos centros en torno a la pendiente media γ_1 . Se asume que los términos U_{0j} y U_{1j} son variables aleatorias con valor esperado cero y varianzas $\sigma_{U_0}^2$ y $\sigma_{U_1}^2$, respectivamente.

También se asume que los términos γ_0 y U_{0j} son independientes entre sí. Y lo mismo vale decir de los términos γ_1 y U_{1j} . Sin embargo, entre los términos β_{0j} y β_{1j} no se asume independencia. La relación entre ambos viene dada por:

$$\rho(\beta_{0j}, \beta_{1j}) = \text{Cov}(\beta_{0j}, \beta_{1j}) / (\sigma_{U_0} \sigma_{U_1})$$

Los gráficos de la Figura 4.3 pueden ayudar a entender el significado de esta relación. Si el tamaño de las medias es independiente del tamaño de las pendientes (es decir, si $\rho(\beta_{0j}, \beta_{1j}) = 0$), se obtienen rectas de regresión como las que muestran los gráficos *a*,

¹ Si la recuperación media de los pacientes es idéntica en todos los centros y la relación entre la recuperación y las puntuaciones basales es la misma en todos los centros, esta ecuación se reduce a la ecuación de regresión lineal para un único centro.

Figura 4.3. Posibles pautas de relación entre X e Y en cinco hipotéticos centros hospitalarios

b , c y f ; en los gráficos a y b todos los centros comparten la misma ecuación de regresión, es decir, $\sigma_{U_0} = \sigma_{U_1} = 0$ en ambos casos (pero con $\gamma_1 = 0$ en a y $\gamma_1 > 0$ en b); en el gráfico c los centros tienen distinta media pero la misma pendiente ($\sigma_{U_0} > 0$, $\sigma_{U_1} = 0$); las rectas del gráfico f indican que los centros difieren tanto en las medias como en las pendientes ($\sigma_{U_0} > 0$, $\sigma_{U_1} > 0$). Si las pendientes de los centros son tanto mayores cuanto mayores son las medias (es decir, si $\rho(\beta_{0j}, \beta_{1j})$ toma un valor positivo) se obtienen rectas como las del gráfico d . Por último, si las pendientes de los centros son tanto menores cuanto mayores son las medias (es decir, si $\rho(\beta_{0j}, \beta_{1j})$ toma un valor negativo) se obtienen rectas como las del gráfico e .

Puesto que tanto las medias (las intersecciones) como la relación entre X e Y (las pendientes) pueden variar de centro a centro, suele resultar útil incluir en el modelo una o más variables del nivel 2 que puedan dar cuenta de esa variabilidad. Por ejemplo, los centros del archivo *Depresión* están clasificados como públicos ($sector = 1$) y privados ($sector = 0$).

Podría darse el caso de que esta diferencia en el nivel 2 fuera responsable (al menos en parte) de la variabilidad existente, no ya solo entre las medias de los centros, sino entre las pendientes que relacionan la recuperación con las puntuaciones basales. Para incluir en el modelo esta variable del nivel 2 podemos hacer

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01} Z_j + U_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} Z_j + U_{1j}\end{aligned}\tag{4.6}$$

(con $Z = \text{sector}$). Llevando a [4.4] los valores de β_{0j} y β_{1j} en [4.6] se obtiene la formulación convencional de un modelo multinivel:

$$Y_{ij} = \gamma_{00} + \gamma_{01}Z_j + U_{0j} + \gamma_{10}x_{ij} + \gamma_{11}x_{ij}Z_j + U_{1j}x_{ij} + E_{ij}$$

Colocando, solo por claridad, los efectos fijos (γ) al principio y los aleatorios (U y E) al final, entre paréntesis, obtenemos

$$Y_{ij} = \gamma_{00} + \gamma_{01}Z_j + \gamma_{10}x_{ij} + \gamma_{11}x_{ij}Z_j + (U_{0j} + U_{1j}x_{ij} + E_{ij}) \quad [4.7]$$

Y haciendo $Y = \text{"recuperación"}$, $x = \text{"cbasal"}$ (puntuaciones basales centradas en la media) y $Z = \text{"sector"}$ (tipo de centro: 1 = "público", 0 = "privado"), tenemos:

- γ_{00} = recuperación media estimada para los pacientes con puntuación basal media ($cbasal = 0$) en los centros privados ($sector = 0$).
- γ_{01} = diferencia entre la recuperación media de los centros públicos ($sector = 1$) y la de los privados ($sector = 0$) en los pacientes con puntuación basal media ($cbasal = 0$).
- γ_{10} = pendiente media (relación entre las puntuaciones basales y la recuperación) en los centros privados ($sector = 0$).
- γ_{11} = diferencia entre las pendientes de los centros públicos y privados.
- U_{0j} = efecto de los centros sobre la recuperación media (variabilidad entre las medias de los centros).
- U_{1j} = efecto del j -ésimo centro sobre la pendiente de los centros privados (variabilidad entre las pendientes de los centros privados).

El modelo propuesto en [4.7] no es un modelo de regresión lineal convencional: no es razonable asumir que los errores son independientes entre sí ni tampoco que la varianza de los errores es la misma en todos los centros. Por un lado, la parte aleatoria del modelo (la parte entre paréntesis) es más compleja que en el modelo de regresión lineal (el cual únicamente incluye E_{ij}); y está claro que los errores no son independientes dentro de cada centro porque los términos U_{0j} y U_{1j} son comunes a todos los sujetos del mismo centro. Por otro lado, no es posible asumir que la varianza de los errores es la misma en todos los centros porque tanto U_{0j} como U_{1j} varían de centro a centro.

La ecuación [4.4] es el **modelo del nivel 1**; la ecuación [4.6] es el **modelo del nivel 2**; la ecuación [4.7] es el **modelo combinado**. El modelo combinado incluye tanto efectos fijos (los que están fuera del paréntesis) como aleatorios (los que están dentro del paréntesis); es, por tanto, un modelo mixto. Los parámetros β son los coeficientes del nivel 1 (los pacientes) y E_{ij} es el término aleatorio del nivel 1. Los parámetros γ son los coeficientes del nivel 2, y U_{0j} y U_{1j} son los términos aleatorios del nivel 2. La varianza de E_{ij} es la varianza del nivel 1; las varianzas de U_{0j} y U_{1j} y sus covarianzas son los componentes de varianza-covarianza del nivel 2.

Con una variable independiente de cada nivel (X del nivel 1 y Z del nivel 2), el modelo [4.7] es un modelo multinivel completo: incluye todos los términos posibles (po-

drían añadirse variables de uno y otro nivel pero esto no cambiaría las características del modelo). Eliminando términos de [4.7] se obtienen el resto de modelos multinivel. En los apartados que siguen se describen, ajustan e interpretan cinco modelos (ver Raudenbush y Brik, 2002, Capítulos 2 y 4), ordenados desde el más simple al más complejo: (1) análisis de varianza de un factor de efectos aleatorios, (2) análisis de regresión con medias como resultados, (3) análisis de covarianza de un factor de efectos aleatorios, (4) análisis de regresión con coeficientes aleatorios y (5) análisis de regresión con medias y pendientes como resultados.

Todos estos modelos se explican utilizando los datos del archivo *Depresión* (puede descargarse de la página web del manual). En concreto, como variables del nivel 1 (los pacientes) utilizaremos dos: *recuperación* (recuperación en la semana 6) y *basal* (puntuaciones en la escala de Hamilton en el momento basal). Como variables del nivel 2 utilizaremos otras dos: *edad* (edad media de los pacientes en cada centro) y *sector* (tipo de centro: público o privado).

Análisis de varianza: un factor de efectos aleatorios

El modelo multinivel más simple posible se obtiene eliminando del modelo [4.7] todo lo relacionado con las variables independientes X y Z . Se obtiene así un modelo mixto sin variables independientes llamado modelo **incondicional** o **nulo**. En el nivel 1 (en el nivel de los pacientes) este modelo adopta la siguiente forma:

$$Y_{ij} = \beta_{0j} + E_{ij} \quad [4.8]$$

En este nivel, la recuperación de los pacientes (Y) se interpreta como el resultado de combinar la recuperación media del centro al que pertenecen (β_{0j}) y los errores o variación aleatoria en torno a esa media (E_{ij}). Se asume que los errores se distribuyen normalmente con media cero y con igual varianza en todos los centros (σ_E^2).

En el nivel 2 (el nivel de los centros), la recuperación media de cada centro (β_{0j}) se interpreta como la combinación de la recuperación media en la población de centros (γ_{00}) y la variación aleatoria de cada centro en torno a esa media (U_{0j}):

$$\beta_{0j} = \gamma_{00} + U_{0j} \quad [4.9]$$

Se asume que el componente aleatorio U_{0j} tiene valor esperado cero y varianza $\sigma_{U_0}^2$. Sustituyendo en [4.8] el valor de β_{0j} en [4.9] se obtiene el modelo mixto multinivel o modelo combinado:

$$Y_{ij} = \gamma_{00} + U_{0j} + E_{ij} \quad [4.10]$$

que no es otra cosa que el modelo de ANOVA de un factor de efectos aleatorios ya estudiado en el capítulo anterior (ver el apartado *Modelo de un factor de efectos aleatorios*), con la única diferencia de que allí no se utilizó esta notación sino otra equivalente más propia de los modelos de ANOVA: $Y_{ij} = \mu + \beta_j + E_{ij}$.

Ejemplo. Análisis de varianza: un factor de efectos aleatorios

Este modelo ya lo hemos ajustado en el capítulo anterior (ver Tablas 3.1 a 3.7) y hemos obtenido las estimaciones que resumen las Tablas 4.1 y 4.2. La Tabla 4.1 contiene una estimación puntual de γ_{00} (*intersección* = 9,15) y un intervalo de confianza para esa estimación (7,06; 11,23). El valor de la intersección (9,15) se refiere a la recuperación media estimada en la población de centros. La tabla también ofrece un estadístico t (se obtiene dividiendo el valor estimado entre su error típico) que permite contrastar la hipótesis nula de que la recuperación media vale cero en la población: puesto que el nivel crítico obtenido (*sig.* < 0,0005) es menor que 0,05, se puede rechazar esa hipótesis nula y afirmar que la recuperación media es mayor que cero.

Tabla 4.1. Estimaciones de los parámetros de efectos fijos

Parámetro	Estimación	Error típico	gl	t	Sig.	Intervalo de confianza 95%	
						Límite inferior	Límite superior
Intersección	9,15	,94	10,30	9,73	,000	7,06	11,23

La Tabla 4.2 ofrece las *estimaciones de los dos parámetros de covarianza* del modelo de un factor: la varianza *entre* los centros (*centro*: $\delta_{U_0}^2 = 9,09$) y la varianza *dentro* de los centros (*residuos*: $\delta_E^2 = 18,00$). La tabla incluye los estadísticos necesarios para contrastar la hipótesis nula de que las correspondientes varianzas poblacionales valen cero. Puesto que en ambos casos el nivel crítico es menor que 0,05, se puede afirmar que ambas varianzas son mayores que cero.

El contraste de la hipótesis relativa a la varianza entre los centros permite valorar el efecto del factor *centro*. El rechazo de esta hipótesis implica que la recuperación media de los pacientes no es la misma en todos los centros. Y dado que el factor analizado es de efectos aleatorios, esta conclusión se refiere a la población de centros de la que han sido seleccionados los 11 incluidos en el análisis.

Tabla 4.2. Estimaciones de los parámetros de covarianza

Parámetro		Estimación	Error típico	Wald Z	Sig.	Intervalo de confianza 95%	
						Límite inferior	Límite superior
Residuos		18,00	1,33	13,57	,000	15,58	20,80
centro	Varianza	9,09	4,28	2,12	,034	3,61	22,89

Las estimaciones de la variabilidad *inter* e *intracentro* que ofrece la Tabla 4.2 están estrechamente relacionadas con el **coeficiente de correlación intraclase** (C_{CI}):

$$C_{CI} = \frac{\delta_{U_0}^2}{\delta_{U_0}^2 + \delta_E^2} \quad [4.11]$$

Este coeficiente indica qué proporción de la varianza total (es decir, de la varianza de la variable dependiente) está explicada por las diferencias entre los centros. También indica el grado de relación o parecido existente entre los pacientes de un mismo centro en comparación con el grado de parecido entre pacientes de centros distintos; por tanto, sirve para valorar si tiene o no sentido utilizar la variable de agrupación (*centro* en nuestro ejemplo) para distinguir entre las unidades del nivel 1 y las del nivel 2, lo cual tiene su importancia si tenemos en cuenta que estamos intentando ajustar modelos multinivel porque estamos contemplando la posibilidad de que el grado de parecido entre pacientes de un mismo centro sea mayor que entre pacientes de centros distintos. En nuestro ejemplo,

$$C_{CI} = 9,09 / (9,09 + 18,00) = 0,34$$

Este resultado indica que las diferencias en la recuperación media de los centros explican el 34% de la variabilidad de la recuperación. O lo que es lo mismo, que tras descontar el efecto de los centros, todavía falta por explicar el 66% de esa variabilidad. También indica que, puesto que aproximadamente un tercio ($C_{CI} = 0,34$) de la variabilidad de la recuperación se debe simplemente al hecho de que los pacientes están agrupados en centros, la modelización multinivel está justificada.

Conviene no olvidar que este modelo *incondicional* o *nulo* sirve de referente para realizar comparaciones con otros modelos más complejos. Según veremos, estas comparaciones se utilizan para evaluar la significación estadística de los términos en que difieren los modelos comparados.

Análisis de regresión: medias como resultados

El modelo nulo (modelo de un factor de efectos aleatorios) estudiado en el apartado anterior ofrece, básicamente, información sobre dos aspectos: la variabilidad dentro de cada centro y la variabilidad entre las medias de los centros. Las diferencias entre los pacientes del mismo centro constituyen la variabilidad del nivel 1. Las diferencias entre las medias de los centros constituyen la variabilidad del nivel 2.

Ambos tipos de variabilidad pueden reducirse utilizando variables independientes del nivel apropiado. Comencemos con la variabilidad del nivel 2. Una vez constatada la existencia de diferencias entre las medias de los centros, el siguiente paso del análisis podría orientarse a indagar si hay alguna variable capaz de dar cuenta de esas diferencias. El archivo *Depresión* incluye una variable que recoge la edad de los pacientes (*edad*), pero no la edad individual de cada paciente, sino la edad media de los pacientes de cada centro (se trata, por tanto, de una variable del nivel 2). Se sabe que la edad está relacionada con el alivio de los síntomas depresivos: éstos tienden a remitir con mayor rapidez en personas jóvenes. Puesto que la edad media de los pacientes no es la misma en todos los centros, las diferencias observadas en la recuperación de los pacientes de distintos centros podrían estar explicadas, al menos en parte, por las diferencias en la edad media de los pacientes.

Respecto del modelo *nulo* presentado en el apartado anterior (ver ecuaciones [4.8] y [4.10]), el modelo de *medias como resultados* únicamente añade una variable independiente medida en el nivel 2. El modelo del nivel 1 no cambia:

$$Y_{ij} = \beta_{0j} + E_{ij} \quad [4.12]$$

Y la variable independiente del nivel 2 interviene en el modelo del nivel 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}z_j + U_{0j} \quad (\text{con } z_j = Z_j - \bar{Z}) \quad [4.13]$$

(en lugar de utilizar las puntuaciones directas, Z , utilizamos las diferenciales o centradas, z , para que la constante γ_{00} tenga un significado claro). Sustituyendo en [4.12] el valor de β_{0j} en [4.13] se obtiene el modelo combinado:

$$Y_{ij} = \gamma_{00} + \gamma_{01}z_j + (U_{0j} + E_{ij}) \quad [4.14]$$

(el paréntesis contiene la parte *aleatoria*). Lo que hace este modelo es pronosticar la recuperación media de cada centro a partir de la edad media de sus pacientes. Puesto que la constante o intersección del nivel 1, β_{0j} (que es la media de la variable dependiente cuando se utilizan variables independientes centradas), es función de coeficientes y variables del nivel 2, a este modelo se le llama modelo de **medias** (o **constantes**, o **intersecciones**) **como resultados**.

A diferencia de lo que ocurre en el modelo nulo, aquí el término U_{0j} no se refiere exactamente al efecto del factor centro, sino al efecto del factor centro tras eliminar el efecto debido a la variable del nivel 2 (z). Del mismo modo, la varianza que recoge la variabilidad entre los centros, $\sigma_{U_0}^2$, ahora es una varianza condicional: indica cómo varían los centros tras eliminar las diferencias atribuibles a la variable z .

Ejemplo. Análisis de regresión: medias como resultados

Este ejemplo muestra cómo ajustar e interpretar un modelo multinivel con una covariable del nivel 2. Vamos a pronosticar el grado de *recuperación* a partir de la edad media (*cedad_media*; recordemos que los valores de esta variable están centrados para que el coeficiente γ_{00} tenga un significado claro):

- En el cuadro de diálogo previo al principal, trasladar la variable *centro* a la lista **Sujetos** y pulsar el botón **Continuar** para acceder al cuadro de diálogo principal.
- Trasladar la variable *recuperación* al cuadro **Variable dependiente** y la variable *cedad_media* (edad media centrada) a la lista **Covariables**.
- Pulsar el botón **Fijos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos fijos* y trasladar la variable *cedad_media* a la lista **Modelo**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Aleatorios** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos aleatorios*, marcar la opción **Incluir intersección** y trasladar la varia-

ble *centro* a la lista **Combinaciones**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

- Pulsar el botón **Estadísticos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Estadísticos* y marcar las opciones **Estimaciones de los parámetros** y **Contrastes sobre los parámetros de covarianza**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestran las Tablas 4.3 y 4.4. La primera de ellas recoge las *estimaciones de los dos parámetros de efectos fijos*: la intersección ($\hat{\gamma}_{00} = 9,54$) y el coeficiente asociado a la variable *cedad_media* ($\hat{\gamma}_{01} = -0,39$). Puesto que la variable *cedad_media* está centrada, el valor de la intersección es la recuperación estimada cuando *edad_media* toma su valor medio (*cedad_media* = 0). Y el valor del coeficiente asociado a la variable *cedad_media* representa la *disminución* estimada en la recuperación (0,39 puntos) por cada año que aumenta la edad media de los pacientes de un centro. Puesto que el nivel crítico asociado a este coeficiente (*sig.* = 0,001) es menor que 0,05, se puede concluir que la edad de los pacientes está relacionada con la recuperación.

Tabla 4.3. Estimaciones de los parámetros de efectos fijos

Parámetro	Estimación	Error típico	gl	t	Sig.
Intersección	9,54	,56	9,55	17,17	,000
cedad_media	-,39	,09	9,59	-4,59	,001

La Tabla 4.4 muestra las *estimaciones de los parámetros de covarianza*. La varianza de los residuos ($\hat{\sigma}_E^2 = 17,99$) es casi idéntica a la obtenida con el modelo *nulo* ($\hat{\sigma}_E^2 = 18,00$; ver Tabla 4.2). Como era de esperar, la variabilidad del nivel 1 no se ha visto afectada por la presencia de una variable del nivel 2. Sin embargo, el valor estimado para la varianza de los centros ($\hat{\sigma}_{U_0}^2 = 2,69$) ha experimentado una reducción muy importante (recordemos que, en el modelo *nulo*, $\hat{\sigma}_{U_0}^2 = 9,09$; ver Tabla 4.2).

Por tanto, la variabilidad del nivel 2 se ha visto afectada por la presencia de una variable del nivel 2. El nivel crítico asociado al estadístico de *Wald* (*sig.* = 0,073) indica que, después de controlar la edad de los pacientes, no parece que los centros difieran en el grado de recuperación. No obstante, dado que el estadístico de *Wald* es muy conservador con muestras pequeñas, quizá sea prudente pensar que todavía queda por explicar parte de las diferencias entre los centros. De hecho, al comparar los estadísticos $-2LL$ asociados a ambos modelos se llega a la conclusión de que la variabilidad entre los centros es significativamente distinta de cero. En concreto, con el modelo *nulo* se obtuvimos $-2LL = 2.199,27$ (ver la Tabla 3.3 del capítulo anterior). Al incluir la covariable *cedad_media* hemos obtenido $-2LL = 2.190,40$ (aunque no se incluye aquí la tabla, el procedimiento ofrece este resultado por defecto). La diferencia entre ambos valores ($2.199,27 - 2.190,40 = 8,87$) se distribuye según ji-cuadrado con 1 grado de libertad,

pues los dos modelos comparados solo difieren en un parámetro: γ_{01} . La probabilidad de encontrar valores iguales o mayores que 8,87 en la distribución ji-cuadrado con 1 grado de libertad vale 0,003. Por tanto, puede concluirse que, después de controlar el efecto de la edad, la recuperación media no es la misma en todos los centros; o, si se prefiere, que la varianza de las medias de los centros es mayor que cero (no es infrecuente encontrar estas inconsistencias entre el estadístico de Wald y la diferencia entre las desviaciones, particularmente con pocos grupos en el segundo nivel).

El coeficiente de correlación intraclase (ver [4.11]) permite precisar qué proporción de la varianza total se debe a diferencias entre los centros:

$$C_{CI} = 2,69 / (2,69 + 17,99) = 0,13$$

Este valor indica que el 13% de la varianza de la variable dependiente todavía es atribuible o puede explicarse por las diferencias entre las medias de los centros. Pero, ahora, este coeficiente es *condicional*: está informando de lo que ocurre con los centros y la recuperación tras controlar el efecto de la edad media.

El C_{CI} asociado al modelo *nulo* valía 0,34. Al incorporar al modelo la variable *cedad_media*, el valor del C_{CI} ha bajado hasta 0,13. Esto es debido a que buena parte de las diferencias observadas entre los centros queda explicada por las diferencias en la edad media de los pacientes. Comparando las estimaciones de los parámetros de covarianza del modelo *nulo* y del modelo que incluye la covariable *cedad_media* es posible conocer la proporción de varianza explicada en el nivel 2: $(9,09 - 2,69) / 9,09 = 0,70$. Es decir, el 70% de las diferencias observadas entre los centros (diferencias en la recuperación media) son diferencias explicadas por la edad media.

Tabla 4.4. Estimaciones de los parámetros de covarianza

Parámetro	Estimación	Error típico	Wald Z	Sig.
Residuos	17,99	1,32	13,58	,000
Intersección [sujeto = centro] Varianza	2,69	1,50	1,79	,073

Análisis de covarianza: un factor de efectos aleatorios

Una covariable del nivel 2 puede ayudar a explicar las diferencias existentes entre las medias de los centros (variabilidad del nivel 2). Pero, puesto que todos los pacientes del mismo centro tienen el mismo valor en una variable del nivel 2 y que la varianza del nivel 1, σ_{ϵ}^2 , se asume que es la misma en todos los centros, es lógico esperar que una variable del nivel 2 no sirva para explicar la variabilidad del nivel 1. Para explicar esta variabilidad (la variabilidad existente entre los pacientes de un mismo centro) es necesario recurrir a variables del nivel 1.

El archivo *Depresión* incluye una variable (*basal*) que recoge las puntuaciones basales de los pacientes. Se sabe que las puntuaciones basales están relacionadas con la

recuperación: ésta tiende a ser mayor cuando las puntuaciones basales son más altas. En consecuencia, las puntuaciones basales de los pacientes podrían ayudar a explicar, al menos en parte, las diferencias observadas entre los pacientes de un mismo centro. Al añadir al modelo de *medias como resultados* (ecuación [4.12]) una variable X del nivel 1, el modelo en ese nivel adopta la forma:

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + E_{ij} \quad (\text{con } x_i = X_i - \bar{X}) \quad [4.15]$$

En el nivel 2, el coeficiente β_{0j} no cambia (ver ecuación [4.13]). Y el coeficiente β_{1j} toma el mismo valor en todos los centros (pues, de momento, solo se están relacionando dos variables del nivel 1):

$$\beta_{1j} = \gamma_{10} \quad [4.16]$$

El coeficiente γ_{10} representa la pendiente media que relaciona la *recuperación* de los pacientes con sus *puntuaciones basales*. Sustituyendo en [4.15] el valor de β_{0j} en [4.13] y el de β_{1j} en [4.16] se obtiene el modelo combinado:

$$Y_{ij} = \gamma_{00} + \gamma_{01}z_j + \gamma_{10}x_{ij} + (U_{0j} + E_{ij}) \quad [4.17]$$

Ejemplo. Análisis de covarianza: un factor de efectos aleatorios

Para ajustar un modelo de estas características basta con repetir los pasos del ejemplo anterior (donde solo se incluye la variable *cedad_media*) añadiendo la variable *cbasal* (puntuaciones en el momento basal centradas) a la lista **Covariables** del cuadro de diálogo principal y a la lista **Modelo** del subcuadro de diálogo *Modelos lineales mixtos: Efectos fijos*. Al añadir esta nueva variable se obtienen, entre otros, los resultados que muestran las Tablas 4.5 y 4.6.

La Tabla 4.5 contiene las *estimaciones de los tres parámetros de efectos fijos* que incluye el modelo: (1) la intersección ($\hat{\gamma}_{00} = 9,51$) es la recuperación estimada para los pacientes con edad media y puntuación basal media (es decir, la recuperación estimada cuando *cedad_media* = 0 y *cbasal* = 0); (2) el coeficiente de regresión asociado a la variable *cedad_media* ($\hat{\gamma}_{01} = -0,34$) toma un valor similar al obtenido antes de incorporar al modelo la variable *cbasal* (ver Tabla 4.3); y (3) el coeficiente asociado a la variable *cbasal* ($\hat{\gamma}_{10} = 0,22$) estima un aumento de 0,22 puntos en la recuperación por cada punto que aumentan las puntuaciones basales. Los tres coeficientes son significativamente distintos de cero (*sig.* < 0,05 en los tres casos).

Tabla 4.5. Estimaciones de los parámetros de efectos fijos

Parámetro	Estimación	Error típico	gl	t	Sig.
Intersección	9,51	,52	9,46	18,46	,000
cedad_media	-,34	,08	9,69	-4,25	,002
cbasal	,22	,03	372,11	6,55	,000

La Tabla 4.6 muestra las *estimaciones de los dos parámetros de covarianza*. El valor estimado para la variabilidad entre los centros ($\hat{\sigma}_{\nu_0}^2$) ha disminuido ligeramente; ha pasado de 2,69 (ver Tabla 4.4) a 2,29. Y la varianza de los residuos ($\hat{\sigma}_E^2$) ha pasado de 18,00 en el modelo nulo (ver Tabla 4.2) a 16,21. Por tanto, al corregir el grado de recuperación mediante las puntuaciones basales, la variabilidad intracentro se ha visto reducida en un 9,9% (pues $100(18,00 - 16,21)/18,00 = 9,9$).

Tabla 4.6. Estimaciones de los parámetros de covarianza

Parámetro	Estimación	Error típico	Wald Z	Sig.
Residuos	16,21	1,20	13,56	,000
Intersección [sujeto = centro] Varianza	2,29	1,30	1,77	,077

Análisis de regresión: coeficientes aleatorios

A los modelos multinivel estudiados hasta ahora se les suele llamar modelos de *constantes o intersecciones aleatorias* porque, en todos ellos, el único coeficiente que varía aleatoriamente de un centro a otro es la intersección β_{0j} . En dos de los modelos estudiados (análisis de varianza de un factor de efectos aleatorios y análisis de regresión con medias como resultados) la pendiente β_{1j} simplemente no existe; y en el tercero (análisis de covarianza de un factor de efectos aleatorios) se le hace tomar un valor fijo.

El modelo de ANCOVA estudiado en el apartado anterior asume que la relación entre la covariable (*cbasal*) y la variable dependiente (*recuperación*) es homogénea en todos los centros ($\beta_{1j} = \gamma_{10}$ para todo j). Sin embargo, para responder correctamente a la cuestión de qué parte de la variabilidad intracentro (variabilidad del nivel 1) puede explicarse por las puntuaciones basales, es decir, para evaluar correctamente la relación existente entre el grado de recuperación y las puntuaciones basales, es necesario obtener *una ecuación de regresión para cada centro* y analizar cómo varían las intersecciones y las pendientes de esas ecuaciones. Al proceder de esta manera se está asumiendo, no solo que los centros pueden diferir en el grado de recuperación (distintas medias), sino que la relación entre el grado de recuperación y las puntuaciones basales puede no ser la misma en todos los centros (distintas pendientes).

Al modelo que recoge este tipo de variación se le llama de **coeficientes aleatorios** justamente porque asume que ambos coeficientes (la intersección y la pendiente) pueden variar de centro a centro. En el nivel 1, el modelo es idéntico al del análisis de covarianza estudiado en el apartado anterior:

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + E_{ij} \quad [4.18]$$

En el nivel 2, el término β_{0j} también se define de idéntica manera en ambos modelos (ver ecuación [4.13]):

$$\beta_{0j} = \gamma_{00} + U_{0j} \quad [4.19]$$

(por supuesto, aquí es posible introducir una o más covariables del nivel 2). La diferencia entre ambos modelos está en la forma de definir la pendiente β_{1j} . En el modelo de análisis de covarianza de un factor aleatorio estudiado en el apartado anterior, el coeficiente β_{1j} se interpreta como una constante (se estima una sola pendiente para todos los centros: γ_{10} ; ver ecuación [4.16]). En el modelo de regresión con coeficientes aleatorios el coeficiente β_{1j} se interpreta como una variable:

$$\beta_{1j} = \gamma_{10} + U_{1j} \quad [4.20]$$

Por tanto, cada centro tiene su propia pendiente (se estiman tantas pendientes como centros). Sustituyendo en [4.18] el valor de β_{0j} en [4.19] y el de β_{1j} en [4.20], el modelo multinivel mixto o combinado queda de la siguiente manera:

$$Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + (U_{0j} + U_{1j} x_{ij} + E_{ij}) \quad [4.21]$$

Y haciendo $Y = \text{“recuperación”}$ y $x = \text{“puntuaciones basales centradas”}$:

- γ_{00} = recuperación media estimada para los pacientes con puntuación basal media (para los pacientes con $cbasal = 0$).
- γ_{10} = pendiente media que relaciona la recuperación con las puntuaciones basales.
- U_{0j} = efecto del j -ésimo centro sobre la recuperación media (variabilidad entre las medias).
- U_{1j} = efecto del j -ésimo centro sobre las pendientes (variabilidad entre las pendientes).

Se asume que los errores del nivel 1, E_{ij} , se distribuyen normalmente con media cero y con la misma varianza σ_E^2 en todos los centros; y que U_{0j} y U_{1j} se distribuyen normalmente con valor esperado cero y varianzas $\sigma_{U_0}^2$ y $\sigma_{U_1}^2$, respectivamente.

Ejemplo. Análisis de regresión: coeficientes aleatorios

Para ajustar e interpretar un modelo de regresión con coeficientes aleatorios utilizando la variable *cbasal* (puntuaciones basales centradas) como variable del nivel 1:

- En el cuadro de diálogo previo al principal, trasladar la variable *centro* (centro hospitalario) a la lista **Sujetos** y pulsar el botón **Continuar** para acceder al cuadro de diálogo principal.
- Trasladar la variable *recuperación* (recuperación en la semana 6) al cuadro **Variable dependiente** y la variable *cbasal* (puntuaciones basales centradas) a la lista **Covariables**.
- Pulsar el botón **Fijos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos fijos* y trasladar la variable *cbasal* a la lista **Modelo**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

- Pulsar el botón **Aleatorios** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos aleatorios*, seleccionar **Sin estructura** en el menú desplegable **Tipo de covarianza**², marcar la opción **Incluir intersección** y trasladar la variable *cbasal* a la lista **Modelo** y la variable *centro* a la lista **Combinaciones**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Estadísticos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Estadísticos* y marcar las opciones **Estimaciones de los parámetros** y **Contrastes sobre los parámetros de covarianza**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestran las Tablas 4.7 y 4.8. La Tabla 4.7 ofrece las estimaciones de los dos *parámetros de efectos fijos* que incluye el modelo que estamos ajustando: (1) la constante o intersección ($\hat{\gamma}_{00} = 9,15$), que sigue siendo una estimación de la recuperación media en la población (para los pacientes con puntuación basal media), y (2) el coeficiente asociado a la variable *cbasal* ($\hat{\gamma}_{10} = 0,37$), que es una estimación de la pendiente media que relaciona las puntuaciones basales con la recuperación. En cada centro se ha estimado una ecuación de regresión relacionando las puntuaciones basales con el grado de recuperación; 0,37 es una estimación de la media de todas esas pendientes. Este valor indica que, por cada punto que aumentan las puntuaciones basales, la ecuación de regresión estima un aumento de 0,37 puntos en la recuperación. El nivel crítico (*sig.* = 0,006) asociado al estadístico *t* permite concluir que la pendiente poblacional media es distinta de cero y, consecuentemente, que las puntuaciones basales están positivamente relacionadas con la recuperación.

Tabla 4.7. Estimaciones de los parámetros de efectos fijos

Parámetro	Estimación	Error típico	gl	t	Sig.
Intersección	9,15	,77	10,60	11,85	,000
cbasal	,37	,11	10,02	3,43	,006

La Tabla 4.8 muestra las cuatro *estimaciones de los parámetros de covarianza* que incluye el modelo: (1) la varianza de los *residuos* ($\hat{\sigma}_E^2$), (2) la varianza de las medias o intersecciones [$NE(1,1) = \hat{\sigma}_{U_0}^2$], (3) la varianza de las pendientes [$NE(2,2) = \hat{\sigma}_{U_1}^2$] y (4) la covarianza entre las medias y las pendientes [$NE(2,1)$]. Las siglas *NE* indican que se ha elegido una matriz **G No Estructurada**. Veamos el significado de cada estimación:

² Los factores de efectos aleatorios imponen una estructura de covarianza a los datos (matriz **G**). En los modelos estudiados hasta ahora se ha utilizado la estructura de covarianza que el SPSS utiliza por defecto: *componentes de la varianza*. Aunque ésta es la estructura de covarianza habitualmente utilizada en los modelos de intersecciones aleatorias, en el modelo de coeficientes aleatorios (en el que no se asume independencia entre los parámetros β_{0j} y β_{1j}) es necesario decidir qué tipo de relación (estructura de covarianza) se desea asignar. Ahora bien, como normalmente no se tiene información sobre esta relación, suele utilizarse un tipo de covarianza *no estructurada*, que equivale a no imponer ningún tipo de estructura predefinida y dejar que sea el procedimiento el que la estime a partir de los datos.

1. La varianza de los *residuos* refleja la variabilidad de la recuperación individual de los pacientes en torno a la recta de regresión de su centro. El valor estimado, 12,64, es menor que el valor estimado con el modelo *nulo* (18,00; ver Tabla 4.2); comparando estas dos estimaciones (la del modelo *nulo* y la del modelo de *coeficientes aleatorios*) es posible saber cuánto disminuye la variabilidad del nivel 1:

$$\text{Reducción en la variabilidad del nivel 1} = (18,00 - 12,64) / 18,00 = 0,30$$

Este resultado indica que, al incluir las puntuaciones basales en el modelo de regresión utilizando una ecuación distinta para cada centro, la variabilidad intracentro se reduce un 30%. Recuérdese que, con una única ecuación de regresión para todos los centros (ver Tabla 4.6), las puntuaciones basales reducían la variabilidad intracentro únicamente un 9,9%.

2. La varianza de las medias o intersecciones ($NE(1,1) = \delta_{U_0}^2 = 6,03$) es mayor que cero ($sig. = 0,034$). Por tanto, puede concluirse que la recuperación media de los centros, es decir, las intersecciones de las ecuaciones de regresión de los distintos centros, no son iguales.
3. La varianza de las pendientes ($NE(2,2) = \delta_{U_1}^2 = 0,11$) es mayor que cero ($sig. = 0,046$). Por tanto, puede concluirse que las pendientes de las ecuaciones de regresión no son iguales en todos los centros; es decir, que la relación entre las puntuaciones basales y el grado de recuperación cambia dependiendo del centro.
4. No existe evidencia de que las pendientes estén relacionadas con las medias ($sig. = 0,448$). Por tanto, la relación intracentro entre las puntuaciones basales y el grado de recuperación no parece ir aumentando o disminuyendo conforme lo hace el tamaño de las medias.

Tabla 4.8. Estimaciones de los parámetros de covarianza (matriz **G**: no estructurada, NE)

Parámetro		Estimación	Error típico	Wald Z	Sig.
Residuos		12,64	,94	13,38	,000
Intersección + cbasal [sujeito = centro]	NE (1,1)	6,03	2,84	2,12	,034
	NE (2,1)	,22	,29	,76	,448
	NE (2,2)	,11	,06	1,99	,046

La ecuación [4.21] incluye cinco parámetros (dos de efectos fijos y tres de efectos aleatorios). Sin embargo, al ajustar el modelo de coeficientes aleatorios se están estimando seis parámetros (dos de efectos fijos y cuatro de efectos aleatorios). El sexto parámetro es la covarianza entre las medias y las pendientes, la cual, al seleccionar una matriz **G** no estructurada se asume que es distinta de cero.

Ahora bien, puesto que la covarianza entre las medias y las pendientes no alcanza la significación estadística ($NE(2,1) = 0,22$; $sig. = 0,448$), puede eliminarse del modelo sin pérdida de ajuste. Cuando no existe evidencia de que las pendientes cambien al cambiar las medias, lo razonable es asumir que la pendiente es la misma en todos los

centros y, consecuentemente con ello, ajustar un modelo eligiendo una matriz **G** que tenga en cuenta esta circunstancia (por ejemplo, *simetría compuesta*).

Eligiendo la opción **Componentes de la varianza** en el subcuadro de diálogo *Modelos lineales mixtos: Efectos aleatorios* (en lugar de la opción **Sin estructura**) se obtienen las estimaciones de los parámetros de covarianza que muestra la Tabla 4.9. El parámetro correspondiente a la covarianza entre las medias y las pendientes ha desaparecido. La varianza de los *residuos* y la varianza de las pendientes (*cbasal [sujeito = centro]*) no se han alterado. Y la varianza de las medias (*intersección [sujeito = centro]*) ha cambiado solo ligeramente.

Tabla 4.9. Estimaciones de los parámetros de covarianza (matriz **G**: componentes de la varianza)

Parámetro	Estimación	Error típico	Wald Z	Sig.
Residuos	12,64	,94	13,38	,000
Intersección [sujeito = centro] Varianza	5,98	2,82	2,12	,034
cbasal [sujeito = centro] Varianza	,11	,06	1,99	,047

Análisis de regresión: medias y pendientes como resultados

Habiendo encontrado que tanto las *medias* (es decir, las intersecciones) como las *pendientes* varían de centro a centro, el siguiente paso lógico consiste en intentar averiguar qué variables podrían dar cuenta de esta variabilidad. Se trata de comprender por qué la recuperación media de unos centros es mayor que la de otros y por qué la relación (la pendiente) entre las puntuaciones basales y la recuperación es mayor en unos centros que en otros. Este es justamente el hecho diferencial o característico de un modelo multinivel: los coeficientes (las medias y las pendientes) del nivel 1 se conciben como *resultados* de los coeficientes y variables del nivel 2.

Al ajustar el modelo de *medias como resultados* hemos encontrado que la edad de los pacientes explica el 70% de las diferencias observadas en la recuperación media de los centros, es decir, el 70% de la variabilidad en las medias. Falta por averiguar qué variable(s) podría(n) dar cuenta de la variabilidad observada en las pendientes que relacionan la recuperación con las puntuaciones basales. El archivo *Depresión* incluye una variable llamada *sector* (tipo de centro hospitalario) con código 1 para los centros públicos y código 0 para los privados. Curiosamente, la relación entre las puntuaciones basales y el grado de recuperación es sensiblemente mayor en los centros públicos ($R_{XY} = 0,79$) que en los privados ($R_{XY} = 0,05$). Se trata, por tanto, de una variable que, en principio, podría ayudar a explicar, al menos en parte, las diferencias encontradas entre las pendientes.

El modelo de regresión que interpreta las medias (intersecciones) y las pendientes como resultados es, en el nivel 1, idéntico al modelo de coeficientes aleatorios:

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + E_{ij} \quad [4.22]$$

Pero en el nivel 2 incluye las variables que se desea utilizar para explicar la variabilidad de las medias y de las pendientes:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}z_j + \gamma_{02}w_j + U_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}z_j + \gamma_{12}w_j + U_{1j}\end{aligned}\quad [4.23]$$

Tanto z como w son variables del nivel 2 (las letras minúsculas indican que se trata de variables *centradas*). Sustituyendo en [4.22] los valores de β_{0j} y β_{1j} en [4.23] tenemos:

$$Y_{ij} = \gamma_{00} + \gamma_{01}z_j + \gamma_{02}w_j + \gamma_{10}x_{ij} + \gamma_{11}x_{ij}z_j + \gamma_{12}x_{ij}w_j + (U_{0j} + U_{1j}x_{ij} + E_{ij}) \quad [4.24]$$

Y haciendo $x = \text{"cbasal"}$ (puntuaciones basales centradas; nivel 1), $z = \text{"cedad_media"}$ (edad media centrada; nivel 2) y $w = \text{"sector"}$ (tipo de centro; nivel 2), el modelo de *medias y pendientes como resultados* propuesto en [4.24] puede formularse como

$$\begin{aligned}Y_{ij} &= \gamma_{00} + \gamma_{01}(\text{cedad_media}) + \gamma_{02}(\text{sector}) + \gamma_{10}(\text{cbasal}) + \gamma_{11}(\text{cbasal})(\text{cedad_media}) \\ &\quad + \gamma_{12}(\text{cbasal})(\text{sector}) + (U_{0j} + U_{1j}(\text{cbasal}) + E_{ij})\end{aligned}$$

Donde:

- γ_{00} = recuperación media cuando las variables *sector*, *cedad_media* y *cbasal* valen cero.
- γ_{01} = efecto de la *edad*; indica cómo cambia la recuperación media de los centros cuando aumenta la edad media entre los pacientes con puntuación basal media (*cbasal* = 0).
- γ_{02} = efecto del *sector*; representa la diferencia en la recuperación media de los centros públicos y privados entre los pacientes con puntuación basal media (*cbasal* = 0).
- γ_{10} = pendiente media que relaciona la recuperación con las puntuaciones basales cuando las variables *sector* y *cedad_media* valen cero.
- U_{0j} = efecto del j -ésimo centro sobre las medias (variabilidad entre las medias).
- U_{1j} = efecto del j -ésimo centro sobre las pendientes (variabilidad entre las pendientes).
- E_{ij} = variabilidad dentro de cada centro (errores aleatorios del nivel 1).

Lo característico de este modelo es que incluye dos interacciones entre variables de distinto nivel: *cbasal* es una variable del nivel 1 (los pacientes); *cedad_media* y *sector* son variables del nivel 2 (los centros):

- γ_{11} = efecto conjunto de las variables *cbasal* y *cedad_media*; indica si la relación entre la recuperación y las puntuaciones basales cambia cuando cambia la edad media de los centros privados (*sector* = 0).
- γ_{12} = efecto conjunto de las variables *cbasal* y *sector*; indica si la relación entre la recuperación y las puntuaciones basales es o no la misma en los centros públicos y en los privados cuando *cedad_media* vale cero.

Se asume que los errores del nivel 1, E_{ij} , se distribuyen normalmente con media cero y con la misma varianza σ_E^2 en todos los centros, y que U_{0j} y U_{1j} se distribuyen normalmente con valor esperado cero y varianzas $\sigma_{U_0}^2$ y $\sigma_{U_1}^2$, respectivamente.

Ejemplo. Análisis de regresión: medias y pendientes como resultados

Para ajustar e interpretar un modelo de regresión que trate las medias y las pendientes como resultados:

- En el cuadro de diálogo previo al principal, trasladar la variable *centro* (centro hospitalario) a la lista **Sujetos** y pulsar el botón **Continuar** para acceder al cuadro de diálogo principal.
- Trasladar la variable *recuperación* (recuperación en la semana 6) al cuadro **Variable dependiente** y las variables *cedad_media* (edad media centrada), *sector* (tipo de centro) y *cbasal* (puntuaciones basales centradas) a la lista **Covariables**.
- Pulsar el botón **Fijos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos fijos* y trasladar a la lista **Modelo** los efectos principales *cedad_media*, *sector* y *cbasal* y las interacciones *cbasal* × *cedad_media* y *cbasal* × *sector*. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Aleatorios** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos aleatorios*, seleccionar **Sin estructura** en el menú desplegable **Tipo de covarianza**, marcar la opción **Incluir intersección**, y trasladar la variable *cbasal* a la lista **Modelo** y la variable *centro* a la lista **Combinaciones**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Estadísticos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Estadísticos* y marcar las opciones **Estimaciones de los parámetros** y **Contrastes sobre los parámetros de covarianza**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestran las Tablas 4.10 y 4.11. La Tabla 4.10 ofrece las estimaciones de los *parámetros de efectos fijos*, que en este modelo son seis: la intersección, los tres efectos principales y las dos interacciones (es decir, todos los coeficientes γ del modelo). Veamos cuál es el significado de cada estimación ayudándonos de los gráficos de la Figura 4.4:

1. La constante o intersección ($\hat{\gamma}_{00} = 8,71$) es una estimación de la recuperación media en la población de centros cuando todas las variables independientes valen cero. El correspondiente nivel crítico (*sig.* < 0,0005) permite afirmar que la recuperación media en la población es distinta de cero.
2. Entre los pacientes con puntuación basal media (*cbasal* = 0), la edad (*cedad_media*) está relacionada negativa ($\hat{\gamma}_{01} = -0,25$) y significativamente (*sig.* = 0,027) con la recuperación. El valor del coeficiente de regresión indica que la recuperación me-

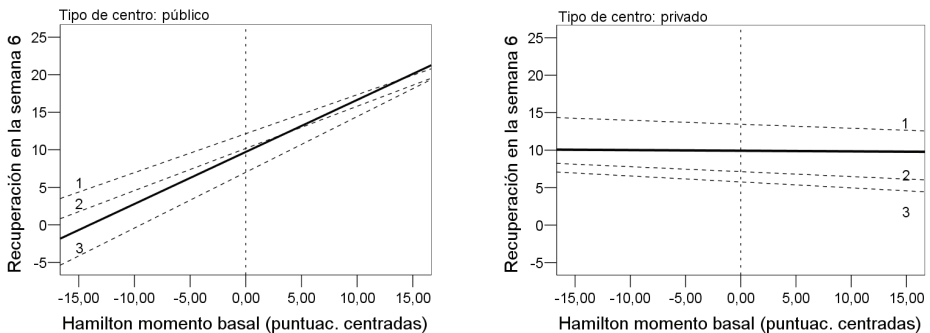
día de los pacientes con puntuación basal media disminuye 0,25 puntos por cada año que aumenta la edad media (en esta interpretación se está asumiendo que la interacción *cedad_media* $\times$ *cbasal* es significativa; como de hecho esa interacción es no significativa, el efecto de la variable *cedad_media* hay que extenderlo a cualquier valor de *cbasal*, no solo a su valor medio). En los gráficos de la Figura 4.4 se puede apreciar este efecto: conforme aumenta la edad (1 = “menos edad”, 3 = “más edad”), las medias o intersecciones (puntos en los que las rectas cortan la línea vertical trazada sobre la puntuación basal cero) son más bajas.

- Entre los pacientes con puntuación basal media, el tipo de centro (*sector*) no parece afectar a la recuperación. El valor del coeficiente ($\hat{\gamma}_{02} = -1,21$) indica que la recuperación estimada para los centros públicos (*sector* = 1) es 1,21 puntos mayor que la estimada para los centros privados (*sector* = 0). Pero esta diferencia no alcanza la significación estadística (*sig.* = 0,343). En los gráficos de la Figura 4.4 puede apreciarse que la recuperación media de los centros públicos y privados es aproximadamente la misma: los puntos de corte de las líneas continuas y más gruesas están aproximadamente a la misma altura ($\hat{\gamma}_{00}$ es una estimación de esa altura).
- No parece que las puntuaciones basales (*cbasal*) estén relacionadas con la recuperación ($\hat{\gamma}_{10} = 0,06$; *sig.* = 0,511). Pero debe tenerse en cuenta que este resultado se

Tabla 4.10. Estimaciones de los parámetros de efectos fijos

Parámetro	Estimación	Error típico	gl	t	Sig.
Intersección	8,71	,88	7,65	9,93	,000
<i>cedad_media</i>	-,25	,09	8,54	-2,67	,027
<i>sector</i>	1,21	1,20	8,27	1,01	,343
<i>cbasal</i>	,06	,09	3,69	,73	,511
<i>cedad_media</i> * <i>cbasal</i>	,00	,01	4,78	-,40	,709
<i>sector</i> * <i>cbasal</i>	,52	,12	4,41	4,30	,010

Figura 4.4 Relación entre las *puntuaciones basales* y la *recuperación* en tres centros públicos (izquierda) y tres privados (derecha). En ambos casos están representados tres centros con edades bajas (1), medias (2) y altas (3). Las líneas continuas son las pendientes medias de cada tipo de centro



refiere a la edad media ($cedad_media = 0$) y a los centros privados ($sector = 0$). En relación con esto, debe prestarse especial atención al comentario del párrafo 6 sobre la interacción entre las variables $sector$ y $cbasal$.

5. El efecto estimado para la interacción entre la edad ($cedad_media$) y las puntuaciones basales ($cbasal$) es nulo; el coeficiente $\hat{\gamma}_{11}$ vale 0 y el nivel crítico 0,709. En los centros privados ($sector = 0$), las pendientes que relacionan las puntuaciones basales y la recuperación no parecen cambiar al cambiar la edad (un coeficiente positivo y significativo asociado a esta interacción estaría indicando que la relación entre las puntuaciones basales y la recuperación es mayor cuanto mayor es la edad).

Ya hemos señalado (párrafo 2) que el aumento en la edad media de los centros va acompañado de una disminución en la recuperación media. Lo que estamos diciendo ahora es que la edad no parece alterar el valor de las pendientes en los centros privados. En los gráficos de la Figura 4.4 se puede apreciar que la relación es muy similar en los tres centros privados: las tres pendientes son prácticamente idénticas (algo parecido ocurre en los centros públicos, pero el coeficiente $\hat{\gamma}_{11}$ se refiere solo a los privados, es decir, a $sector = 0$).

6. En relación con la interacción entre el tipo de centro ($sector$) y las puntuaciones basales ($cbasal$), el coeficiente $\hat{\gamma}_{12}$ toma un valor positivo (0,52) y tiene asociado un nivel crítico menor que 0,05 ($sig. = 0,010$). Por tanto, cuando $cedad_media$ vale cero, la pendiente que relaciona la recuperación y las puntuaciones basales no es la misma en los centros públicos y en los privados: la pendiente media en los centros públicos ($sector = 1$) es 0,52 puntos mayor que en los privados ($sector = 0$). Es decir, la relación entre las puntuaciones basales y la recuperación es significativamente mayor en los centros públicos que en los privados.

Por tanto, aunque el resultado del párrafo 4 indica que, en los centros privados, las puntuaciones basales no están relacionadas con la recuperación, parece que no es eso lo que ocurre en los centros públicos. Precisamente el hecho más llamativo de los gráficos de la Figura 4.4 es que, mientras la pendiente media (línea continua) de los centros públicos es alta y positiva, la pendiente media de los centros privados es prácticamente nula. El coeficiente $\hat{\gamma}_{12} = 0,52$ refleja justamente esta diferencia entre las pendientes medias.

Finalmente, la Tabla 4.11 ofrece las *estimaciones de los parámetros de covarianza*, que en este modelo son cuatro: (1) la varianza de los residuos ($residuos = \delta_E^2$), (2) la varianza de las medias o intersecciones [$NE(1,1) = \delta_{U_0}^2$], (3) la varianza de las pendientes [$NE(2,2) = \delta_{U_1}^2$] y (4) la covarianza entre las medias y las pendientes [$NE(2,1)$]. Veamos el significado de cada estimación:

1. La varianza de los residuos refleja la variabilidad de la recuperación individual de los pacientes en torno a la recta de regresión de sus respectivos centros. El valor estimado para esta variabilidad (12,73) es muy parecido al estimado con el modelo de coeficientes aleatorios del ejemplo anterior (tal como cabía esperar, las covariables del nivel 2 no contribuyen a reducirlo).

2. La varianza de las medias es sensiblemente menor que la obtenida con el modelo de coeficientes aleatorios (3,45 frente a 6,03; ver Tabla 4.8); al incorporar las variables *cedad_media* y *sector*, la varianza de las medias de los centros (es decir, la variabilidad del nivel 2) se reduce un 42,8% (pues $100(6,03 - 3,45)/6,03 = 42,8$). Esto equivale a afirmar que, tras eliminar de la recuperación el efecto atribuible a las puntuaciones basales, las covariables *cedad_media* y *sector* explican el 42,8% de las diferencias entre los centros (al interpretar este porcentaje debe tenerse en cuenta que si las diferencias entre centros fueran pequeñas, la varianza *explicable* también lo sería, y un alto porcentaje de reducción de esa varianza seguiría siendo una cantidad pequeña). Por supuesto, como el efecto de *cedad_media* es estadísticamente significativo (*sig.* = 0,027) y el de *sector* no lo es (*sig.* = 0,343), cabe suponer que la mayor parte de ese 42,8% de reducción de las diferencias entre centros corresponde a la edad media. De hecho, cuando no se tienen en cuenta otras variables, la edad media, ella sola, consigue reducir un 70% la variabilidad entre las medias de los centros (ver Tabla 4.4).

El nivel crítico asociado al estadístico de *Wald* (*sig.* = 0,074) no permite rechazar la hipótesis nula de que la varianza poblacional de las medias de los centros vale cero, es decir, no permite rechazar la hipótesis nula de que la recuperación media es la misma en todos los centros. Por tanto, cuando se controla el efecto de la *edad*, el del *tipo de centro* y el de las *puntuaciones basales*, las diferencias en la recuperación media de los centros se reducen lo bastante como para dejar de ser estadísticamente significativas.

3. La varianza de las pendientes, que los resultados del modelo anterior nos llevaron a concluir que era distinta de cero, ha dejado de ser estadísticamente significativa (*sig.* = 0,297). Por tanto, una vez controlado el efecto de las covariables *cedad_media* y *sector*, parece que las diferencias entre las pendientes de los distintos centros desaparecen. Y teniendo en cuenta lo que ocurre con las estimaciones de los efectos fijos, cabe suponer que las diferencias entre las pendientes han desaparecido al controlar el efecto de la covariable *sector*.
4. Por último, al igual que ocurría en el modelo de coeficientes aleatorios del ejemplo anterior, tampoco ahora existe evidencia de que las medias estén relacionadas con las pendientes (*sig.* = 0,917); por tanto, no puede afirmarse que la relación intra-centro entre las puntuaciones basales y la recuperación aumente o disminuya en función del tamaño de las medias (el valor estimado para la covarianza entre las medias y las pendientes es -0,01).

Tabla 4.11. Estimaciones de los parámetros de covarianza (matriz **G**: no estructurada, NE)

Parámetro		Estimación	Error típico	Wald Z	Sig.
Residuos		12,73	,96	13,25	,000
Intersección + cbasal [sujeito = centro]	NE (1,1)	3,45	1,93	1,78	,074
	NE (2,1)	-,01	,14	-,10	,917
	NE (2,2)	,03	,03	1,04	,297

El hecho de que la varianza de las pendientes no sea significativamente distinta de cero está indicando que la mayor parte de la variabilidad entre las pendientes (variabilidad detectada con el modelo de coeficientes aleatorios; ver Tabla 4.8) está explicada por las covariables incluidas en el análisis. Pero también está indicando que el parámetro que recoge la variabilidad entre las pendientes (γ_{11}) puede ser excluido del modelo sin pérdida de ajuste. De hecho, en el modelo actual, el estadístico $-2LL$ vale 2.085,54 (aunque no se incluye aquí la tabla con los estadísticos de ajuste global, el procedimiento la ofrece por defecto). Y eliminando γ_{11} (para esto basta con quitar la variable *cbasal* de la lista **Modelo** en el subcuadro de diálogo *Modelos lineales mixtos: Efectos aleatorios*), se obtiene un valor de 2.088,13. La diferencia entre ambos modelos es de 2 parámetros (el coeficiente γ_{11} y el referido a la covarianza entre las medias y las pendientes, que desaparece al eliminar γ_{11}) y la diferencia entre los respectivos estadísticos $-2LL$ es de $2.088,13 - 2.085,54 = 2,59$ puntos. La probabilidad de encontrar valores mayores que 2,59 en la distribución ji-cuadrado con 2 grados de libertad vale 0,274. Puesto que este valor es mayor que 0,05, podemos concluir que no existe evidencia de que el coeficiente γ_{11} sea distinto de cero. Por tanto, no parece que incluyendo pendientes aleatorias se consiga mejor ajuste que no incluyéndolas. Y a igual ajuste, lo razonable es optar por el modelo más simple.

Al eliminar *cbasal* de la lista **Modelo** en el subcuadro de diálogo *Modelos lineales mixtos: Efectos aleatorios* se obtienen las estimaciones de los *parámetros de covarianza* que muestra la Tabla 4.12: han desaparecido las estimaciones correspondientes a la varianza de las pendientes y a la covarianza entre las medias y las pendientes.

Tabla 4.12. Estimaciones de los parámetros de covarianza

Parámetro	Estimación	Error típico	Wald Z	Sig.
Residuos	13,17	,97	13,52	,000
Intersección [sujeto = centro] Varianza	3,20	1,79	1,79	,074

Curvas de crecimiento

Al hablar de estructuras multinivel se tiende a pensar, antes que nada, en objetos o sujetos individuales agrupados en contextos físicos o sociales de mayor orden: individuos agrupados en familias, estudiantes agrupados en colegios, etc. Esto es justamente lo que hemos hecho nosotros al trabajar con pacientes agrupados en centros. Sin embargo, las estructuras multinivel también se dan cuando varias observaciones están anidadas en una unidad de análisis más amplia.

Las medidas repetidas, por ejemplo, pueden considerarse anidadas en los sujetos del mismo modo que los estudiantes en los colegios o los pacientes en los hospitales y, en ese sentido, constituyen una estructura jerárquica que puede ser abordada desde la perspectiva de la modelización multinivel: con las medidas repetidas de un mismo sujeto puede obtenerse una ecuación de regresión (una ecuación por sujeto) de la misma mane-

ra que con los pacientes de un mismo hospital (una ecuación por hospital). A las ecuaciones basadas en medidas repetidas se les llama *curvas de crecimiento* y suelen utilizarse para valorar el cambio individual (ver Raudenbush, 2001; Singer y Willett, 2003).

Medidas repetidas: coeficientes aleatorios

Comencemos formulando un modelo de **coeficientes aleatorios** para las medidas repetidas del archivo *Depresión repetidas multinivel* (los datos están dispuestos tal como exige el procedimiento **Mixed**). En el nivel 1, este modelo adopta la forma:

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + E_{ij} \quad [4.25]$$

Y en el nivel 2:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + U_{0j} \\ \beta_{1j} &= \gamma_{10} + U_{1j} \end{aligned} \quad [4.26]$$

Por tanto, cada caso (en nuestro ejemplo, cada paciente) tiene su propia intersección y su propia pendiente: se estiman tantas ecuaciones de regresión (tantas curvas de crecimiento) como casos tiene el archivo. Sustituyendo en [4.25] los valores de β_{0j} y β_{1j} en [4.26], el modelo multinivel combinado queda de la siguiente manera:

$$Y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + (U_{0j} + U_{1j}x_{ij} + E_{ij}) \quad [4.27]$$

Las unidades del nivel 1 son cada una de las medidas repetidas; las unidades del nivel 2 son los casos (pacientes). Tomando *cmomento* (momento centrado en la semana 6)³ como variable independiente del nivel 1, el modelo de *coeficientes aleatorios* propuesto en [4.27] queda de la siguiente manera:

$$Y_{ij} = \gamma_{00} + \gamma_{10}(\text{cmomento}) + U_{0j} + U_{1j}(\text{cmomento}) + E_{ij}$$

Este modelo multinivel intenta explicar las puntuaciones *hamilton* (Y) a partir de:

- γ_{00} = media de la variable dependiente (*hamilton*) cuando la variable *cmomento* vale cero (es decir, en la sexta semana).
- γ_{10} = pendiente media que expresa la relación entre el paso del tiempo (*cmomento*) y la variable dependiente *hamilton*.

³ Centrar la variable *momento* en la semana 6 tiene el objetivo de referir las medias y las pendientes del modelo al momento en el que los pacientes finalizan el tratamiento, no al momento en el que lo inician. De este modo, tanto las medias como las pendientes tienen un significado más claro. Por ejemplo, la intersección es la media de la variable dependiente cuando la variable independiente vale cero; puesto que la variable independiente representa los momentos en los que se han realizado las mediciones, si se asigna el código cero al momento basal, la intersección será la media de las puntuaciones basales; si se asigna el valor cero a la media de todos los momentos, la intersección será la media de la variable dependiente cuando la variable independiente toma su valor medio; si se asigna un cero a la última medición (como en nuestra variable *cmomento*), la intersección será una estimación de la media de la variable dependiente al final del tratamiento.

U_{0j} = efecto de los casos (los pacientes) o variabilidad de las medias de los casos en torno a γ_{00} (la media total en la sexta semana).

U_{1j} = variabilidad de las pendientes (una por cada caso) en torno a la pendiente media γ_{10} .

Los E_{ij} siguen siendo los errores aleatorios del primer nivel; representan la variabilidad intrasujetos o variabilidad de las distintas puntuaciones del mismo caso en torno a su puntuación media. Nótese que el modelo no incluye covariables del nivel 2 (aunque podría hacerlo) y que la presencia de U_{0j} y U_{1j} indica que tanto las intersecciones como las pendientes se están considerando aleatorias.

La información que ofrece un modelo de coeficientes aleatorios aplicado a un diseño de medidas repetidas (con una única variable independiente del nivel 1) es exactamente la misma que la de un ANOVA de un factor de medidas repetidas más un par de detalles: (1) la variabilidad de las pendientes y (2) la relación entre las intersecciones y las pendientes (ver Cnaan, Laird y Slasor, 1997).

Ejemplo. Medidas repetidas: coeficientes aleatorios

Para ajustar un modelo de regresión de coeficientes aleatorios propuesto en [4.27] a los datos del archivo *Depresión repetidas multinivel* (el archivo puede descargarse de la página web del manual):

- En el cuadro de diálogo previo al principal, trasladar la variable *id* (identificación de caso) a la lista **Sujetos** y pulsar el botón **Continuar** para acceder al cuadro de diálogo principal.
- Trasladar la variable *hamilton* (puntuaciones escala Hamilton) al cuadro **Variable dependiente** y la variable *cmomento* (momento centrado en la semana 6) a la lista **Covariables**. La variable *cmomento* puede tratarse como una covariable porque es una variable cuantitativa: sus códigos reflejan el número de semanas transcurridas entre mediciones; por otro lado, recordemos que la variable *cmomento* se ha centrado en la semana 6 para que la intersección γ_{00} quede referida a ese momento.
- Pulsar el botón **Fijos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos fijos* y trasladar la variable *cmomento* a la lista **Modelo**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Aleatorios** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos aleatorios*, seleccionar **Sin estructura** en el menú desplegable **Tipo de covarianza**, marcar la opción **Incluir intersección** y trasladar la variable *cmomento* a la lista **Modelo** y la variable *id* a la lista **Combinaciones**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Estadísticos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Estadísticos* y marcar las opciones **Estimaciones de los parámetros** y **Contrastes sobre los parámetros de covarianza**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestran las Tablas 4.13 y 4.14.

La Tabla 4.13 ofrece las *estimaciones de los dos parámetros de efectos fijos*: (1) la constante o *intersección* ($\hat{\gamma}_{00} = 15,23$) de la ecuación de regresión de *cmomento* sobre *hamilton* (que es una estimación de la media de la variable dependiente *hamilton* cuando *cmomento* vale cero, es decir, en la sexta semana), y (2) el coeficiente asociado a la variable *cmomento* ($\hat{\gamma}_{10} = -1,62$) en esa misma ecuación de regresión (que es una estimación de la pendiente media). El nivel crítico (*sig.* < 0,0005) asociado a la variable *cmomento* permite rechazar la hipótesis de relación nula y afirmar que el coeficiente es distinto de cero; por tanto, puede concluirse que *cmomento* (el paso del tiempo) está negativamente relacionado con las puntuaciones *hamilton*. El valor del coeficiente indica que las puntuaciones *hamilton* disminuyen, en promedio, 1,62 puntos con cada semana de tratamiento.

Tabla 4.13. Estimaciones de los parámetros de efectos fijos

Parámetro	Estimación	Error típico	gl	t	Sig.
Intersección	15,23	,35	378	43,66	,000
cmomento	-1,62	,05	378	-35,25	,000

La Tabla 4.14 muestra las *estimaciones de los cuatro parámetros de covarianza* que incluye el modelo propuesto: (1) la varianza del primer nivel (*residuos*); (2) la varianza de las medias o intersecciones [*NE*(1,1)]; (3) la covarianza entre las medias y las pendientes [*NE*(2,1)]; y (4) la varianza de las pendientes [*NE*(2,2)]:

1. La varianza de los *residuos* (2,30) refleja en qué grado varían las puntuaciones (las medidas repetidas) de cada paciente. Esta varianza representa la variabilidad del primer nivel; y el correspondiente nivel crítico (*sig.* < 0,0005) permite concluir que es distinta de cero.
2. El procedimiento calcula 379 ecuaciones de regresión (una por paciente) relacionando *cmomento* (el tiempo medido en semanas) con *hamilton* (las puntuaciones en la escala Hamilton). La varianza de las medias o intersecciones (44,53) es una estimación de la variabilidad existente entre las medias o intersecciones de esas 379 ecuaciones. Puesto que esta varianza es distinta de cero (*sig.* < 0,0005), se puede concluir que las medias de los pacientes en la variable *hamilton* en la sexta semana (cuando *cmomento* = 0) no son iguales.
3. la covarianza entre las medias o intersecciones y las pendientes indica si existe relación entre el tamaño de las medias (puntuación media de cada paciente) y el de las pendientes (relación entre *cmomento* y *hamilton* en cada paciente). El valor de esta covarianza es positivo (2,65) y distinto de cero (*sig.* < 0,0005). Por tanto, se puede concluir que la relación entre el paso del tiempo (*cmomento*) y las puntuaciones en la escala Hamilton (*hamilton*) es tanto mayor cuanto mayores son las puntuaciones medias de los pacientes.

4. La varianza de las pendientes indica cómo varían las pendientes individuales (una por paciente) en torno a la pendiente media de todos los pacientes. Esta varianza vale 0,68 y es distinta de cero (*sig.* < 0,0005). Por tanto, se puede concluir que la pendiente que relaciona el paso del tiempo (*cmomento*) y las puntuaciones en la escala Hamilton (*hamilton*) no es la misma en todos los pacientes.

Tabla 4.14. Estimaciones de los parámetros de covarianza

Parámetro		Estimación	Error típico	Wald Z	Sig.
Residuos		2,30	,12	19,47	,000
Intersección + cmomento [sueto = id]	NE (1,1)	44,53	3,36	13,26	,000
	NE (2,1)	2,65	,35	7,59	,000
	NE (2,2)	,68	,06	11,71	,000

Medidas repetidas: medias y pendientes como resultados

Puesto que tanto las medias como las pendientes varían de paciente a paciente, el siguiente paso del análisis debería ir dirigido a averiguar qué variables podrían dar cuenta de esa variabilidad. Se trata de intentar comprender por qué unos pacientes tienen puntuaciones más altas que otros y por qué la relación (la pendiente) entre el momento de la medición y las puntuaciones *hamilton* es más alta en unos pacientes que en otros. Esto requiere ajustar un modelo multinivel con **medias y pendientes como resultados**.

En el nivel 1, este modelo adopta la misma forma que el modelo de coeficientes aleatorios (ver ecuación [4.25]). Pero, en el nivel 2,

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}z_j + \gamma_{02}w_j + U_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}z_j + \gamma_{12}w_j + U_{1j}\end{aligned}\quad [4.28]$$

Y sustituyendo en [4.25] los valores de β_{0j} y β_{1j} en [4.28], el modelo multinivel mixto o combinado queda de la siguiente manera:

$$Y_{ij} = \gamma_{00} + \gamma_{01}z_j + \gamma_{02}w_j + \gamma_{10}x_{ij} + \gamma_{11}x_{ij}z_j + \gamma_{12}x_{ij}w_j + (U_{0j} + U_{1j}x_{ij} + E_{ij}) \quad [4.29]$$

Ahora, x es una variable del nivel 1 (las medidas repetidas), y w y z son variables del nivel 2 (los casos). Con este modelo se pretende averiguar si las variables w y z ayudan a explicar la variabilidad observada en las medias y en las pendientes.

Haciendo x = “cmomento” (momento centrado en la semana 6), z = “tto” (tratamiento: estándar, combinado) y w = “cbasal” (puntuaciones basales centradas), el modelo de *medias y pendientes como resultados* propuesto en [4.29] puede formularse como:

$$\begin{aligned}Y_{ij} &= \gamma_{00} + \gamma_{01}(\text{tto}) + \gamma_{02}(\text{cbasal}) + \gamma_{10}(\text{cmomento}) + \gamma_{11}(\text{cmomento})(\text{tto}) + \\ &\quad + \gamma_{12}(\text{cmomento})(\text{cbasal}) + (U_{0j} + U_{1j}(\text{cmomento}) + E_{ij})\end{aligned}$$

Este modelo multinivel intenta explicar las puntuaciones *hamilton* (Y_{ij}) a partir de:

- γ_{00} = media de la variable dependiente (*hamilton*) cuando *cbasal* y *cmomento* valen cero y cuando se aplica el tratamiento combinado (que es el nivel de la variable *tto* que el procedimiento fijará en cero; ver Tabla 4.15).
- γ_{01} = efecto principal de *tto*; refleja la diferencia entre las medias de los pacientes que reciben el tratamiento estándar y los que reciben el tratamiento combinado, pero solamente en la sexta semana, que es cuando la variable con la que interaccionan los tratamientos (*cmomento*) vale cero.
- γ_{02} = efecto principal de *cbasal*; indica cómo cambian las puntuaciones *hamilton* al aumentar *cbasal*, pero solamente en la sexta semana, que es cuando la variable con la que interaccionan las puntuaciones basales (*cmomento*) vale cero.
- γ_{10} = efecto principal de la variable *cmomento* (pendiente que relaciona *cmomento* con *hamilton*) cuando *tto* y *cbasal* (las dos variables con las que interacciona *cmomento*) valen cero; indica si, entre los pacientes con puntuación basal media que reciben el tratamiento combinado, las puntuaciones *hamilton* cambian con el momento de la medición.
- γ_{11} = efecto conjunto de las variables *cmomento* y *tto*; indica si el efecto de los tratamientos es o no el mismo (se mantiene constante) en los diferentes momentos; este coeficiente permite contrastar la hipótesis nula relativa al efecto de la interacción entre *cmomento* y *tto*.
- γ_{12} = efecto conjunto de las covariables *cmomento* y *cbasal*; indica si la relación entre el paso del tiempo (*cmomento*) y la variable dependiente (*hamilton*) cambia cuando aumentan las puntuaciones basales; permite contrastar la hipótesis nula relativa al efecto de la interacción entre *cmomento* y *cbasal*.
- U_{0j} = efecto del *j*-ésimo sujeto sobre la variable dependiente *hamilton*; indica cómo varía la media de cada sujeto respecto de la media total.
- U_{1j} = efecto del *j*-ésimo sujeto sobre las pendientes que relacionan la variable *cmomento* con la variable dependiente *hamilton*; indica cómo varía la pendiente de cada sujeto respecto de la pendiente media.

Los E_{ij} siguen siendo los errores aleatorios del nivel 1, los cuales se asume que se distribuyen normalmente con media cero y con igual varianza en todos los sujetos. También se asume que U_{0j} y U_{1j} se distribuyen normalmente.

Ejemplo. Medidas repetidas: medias y pendientes como resultados

Para ajustar un modelo de medias y pendientes como resultados a los datos del archivo *Depresión repetidas multinivel*:

- Utilizar la opción **Seleccionar casos** del menú **Datos** para filtrar las puntuaciones de las semanas 2, 4 y 6, dejando fuera la semana 0 o momento basal (esto es necesario

hacerlo porque el momento basal se va a incluir en el análisis como covariable); filtrar también los tratamientos 1 y 2 (*estándar y combinado*) dejando fuera el 3 (*otro*); la finalidad de esto último es facilitar la interpretación de los coeficientes asociados a la variable *tto*.

- En el cuadro de diálogo previo al principal, trasladar la variable *id* (identificación de caso) a la lista **Sujetos** y pulsar el botón **Continuar** para acceder al cuadro de diálogo principal.
- Trasladar la variable *hamilton* (puntuaciones Hamilton) al cuadro **Variable dependiente**, la variable *tto* (tratamiento) a la lista **Factores** y las variables *cbasal* (puntuaciones Hamilton momento basal centradas) y *cmomento* (momento de la medición centrado en la semana 6) a la lista **Covariables**.
- Pulsar el botón **Fijos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos fijos* y trasladar a la lista **Modelo** los efectos principales *tto*, *cmomento* y *cbasal*, y las interacciones *tto* × *cmomento* y *cbasal* × *cmomento*. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Aleatorios** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Efectos aleatorios*, seleccionar **Sin estructura** en el menú desplegable **Tipo de covarianza**, marcar la opción **Incluir intersección** y trasladar la variable *cmomento* a la lista **Modelo** y la variable *id* a la lista **Combinaciones**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Estadísticos** para acceder al subcuadro de diálogo *Modelos lineales mixtos: Estadísticos* y marcar las opciones **Estimaciones de los parámetros** y **Contrastes sobre los parámetros de covarianza**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestran las Tablas 4.15 a 4.17.

Las Tablas 4.15 y 4.16 contienen información relativa a los *efectos fijos*, que en el modelo que estamos ajustando son seis: la constante o intersección, los tres efectos principales y las dos interacciones (es decir, los coeficientes gamma de la ecuación [4.29]). Cada estadístico *F* de la Tabla 4.15 permite contrastar la hipótesis de que el correspondiente efecto es nulo. Los resultados de la tabla indican que todos los efectos fijos son distintos de cero (*sig.* < 0,0005 en todos los casos).

Tabla 4.15. Contraste de los efectos fijos

Origen	Numerador df	Denominador df	Valor F	Sig.
Intersección	1	322	2966,87	,000
tto	1	322	47,20	,000
cbasal	1	322	264,30	,000
cmomento	1	322	1217,90	,000
cmomento(tto)	1	322	31,90	,000
cbasal * cmomento	1	322	14,48	,000

La Tabla 4.16 ofrece las estimaciones de los *parámetros de efectos fijos*. El modelo propuesto incluye 6 de estos parámetros (los 6 coeficientes gamma que incluye la ecuación [4.29]: (1) la constante o intersección, γ_{00} ; (2) el efecto de los tratamientos, γ_{01} ; (3) el efecto de las puntuaciones basales, γ_{02} ; (4) el efecto del paso del tiempo, γ_{10} ; (5) el efecto de la interacción entre los tratamientos y el paso del tiempo, γ_{11} ; y (6) el efecto de la interacción entre las puntuaciones basales y el paso del tiempo, γ_{12} :

1. Puesto que las covariables *cbasal* y *cmomento* están centradas (en la media *cbasal* y en la sexta semana *cmomento*) y el tratamiento que el procedimiento fija en cero es el combinado (*tto* = 2), la intersección ($\hat{\gamma}_{00} = 13,60$) es una estimación de la media de la variable *hamilton* a las seis semanas de tratamiento para los pacientes con puntuación basal media que han recibido el tratamiento combinado. El nivel crítico asociado al estadístico *t* (*sig.* < 0,0005) indica que el valor poblacional de esa media es distinto de cero.
2. La media de las puntuaciones *hamilton* en la sexta semana es $\hat{\gamma}_{01} = 3,94$ puntos mayor con el tratamiento estándar que con el combinado (*sig.* < 0,0005). El resultado está referido a la sexta semana, que es el momento en el que la variable con la que interaccionan los tratamientos (*cmomento*) vale cero.
3. Las puntuaciones basales (*cbasal*) se relacionan positiva ($\hat{\gamma}_{02} = 0,67$) y significativamente (*sig.* < 0,0005) con las puntuaciones en la variable *hamilton*. La ecuación de regresión estima que, en la sexta semana (es decir, cuando *cmomento* = 0), por cada punto que aumentan las puntuaciones basales (*cbasal*) las puntuaciones *hamilton* aumentan, en promedio, 0,67 puntos.
4. El factor intrasujetos *cmomento* está relacionado negativa ($\hat{\gamma}_{10} = -1,79$) y significativamente (*sig.* < 0,0005) con las puntuaciones *hamilton*. La ecuación de regresión estima que, entre los pacientes con puntuación basal media (*cbasal* = 0) que reciben el tratamiento combinado (*tto* = 2), las puntuaciones *hamilton* disminuyen, en promedio, 1,79 puntos con cada semana de tratamiento.
5. El coeficiente relativo a la interacción entre el paso del tiempo (*cmomento*) y los tratamientos (*tto*) toma un valor positivo ($\hat{\gamma}_{11} = 0,50$) y significativo (*sig.* < 0,0005). Este resultado permite concluir que, una vez controlado el efecto de las puntuacio-

Tabla 4.16. Estimaciones de los parámetros de efectos fijos

Parámetro	Estimación	Error típico	gl	t	Sig.
Intersección	13,60	,33	322	40,62	,000
[tto=1]	3,94	,57	322	6,87	,000
[tto=2]	0 ^a	0	.	.	.
cbasal	,67	,04	322	16,26	,000
cmomento	-1,79	,05	322	-34,62	,000
cmomento([tto=1])	,50	,09	322	5,65	,000
cmomento([tto=2])	0 ^a	0	.	.	.
cbasal * cmomento	-,02	,01	322	-3,81	,000

a. Se ha establecido este parámetro en cero porque es redundante.

nes basales, la relación entre el paso del tiempo (*cmomento*) y las puntuaciones *hamilton* cambia en función del tratamiento: la pendiente media entre los pacientes que reciben el tratamiento estándar (*tto* = 1) es 0,50 puntos mayor que entre los que reciben el tratamiento combinado (*tto* = 2).

6. El coeficiente asociado al efecto de la interacción entre el paso del tiempo (*cmomento*) y las puntuaciones basales (*cbasal*) toma un valor negativo ($\gamma_{12} = -0,02$) y significativo (*sig.* < 0,0005). Esto significa que la relación entre el paso del tiempo (*cmomento*) y las puntuaciones *hamilton* (que sabemos que es de tendencia negativa por el párrafo 4), es tanto más negativa cuanto mayores son las puntuaciones basales. El valor del coeficiente (-0,02) indica que por cada punto que aumentan las intersecciones de las ecuaciones que relacionan *cmomento* con *hamilton*, la pendiente de esas ecuaciones (que es negativa) disminuye 0,02 puntos.

Ya hemos señalado (párrafo 3) que las puntuaciones basales afectan a las medias (a las intersecciones). Lo que estamos afirmando ahora es que las puntuaciones basales también afectan a las pendientes.

La Tabla 4.17 recoge las *estimaciones de los cuatro parámetros de covarianza* que incluye el modelo propuesto: (1) la varianza de los residuos (*residuos*); (2) la varianza de las medias o intersecciones [*NE(1,1)*]; (3) la covarianza entre las medias y las pendientes [*NE(2,1)*]; y (4) la varianza de las pendientes [*NE(2,2)*]:

1. La varianza de los residuos refleja la variabilidad entre las medidas repetidas de cada paciente. Es la variabilidad del primer nivel. Su valor estimado (1,60) es significativamente distinto de cero (*sig.* < 0,0005).
2. La varianza de las intersecciones (22,54), es decir la variabilidad entre las medias de los pacientes, es sensiblemente menor que la obtenida con el modelo de coeficientes aleatorios (44,53; ver Tabla 4.14). La incorporación de las variables *tto* y *cbasal* ha hecho que esta varianza se reduzca a la mitad, aunque sigue siendo significativamente distinta de cero (*sig.* < 0,0005).
3. Las intersecciones siguen relacionadas positiva y significativamente con las pendientes. El valor estimado para la *covarianza* (2,60; *Sig.* < 0,0005) indica que, una vez controlado el efecto de las variables *tto* y *cbasal*, la relación intrapaciente entre *cmomento* y *hamilton* va aumentando conforme lo hace el valor de las intersecciones (esto no ha cambiado respecto de lo que ocurría antes de incluir las variables *tto* y *cbasal*; ver Tabla 4.14).

Tabla 4.17. Estimaciones de los parámetros de covarianza

Parámetro		Estimación	Error típico	Wald Z	Sig.
Residuos		1,60	,13	12,75	,000
Intersección + cmomento [sujeito = id]	NE (1,1)	22,54	1,88	11,96	,000
	NE (2,1)	2,60	,27	9,76	,000
	NE (2,2)	,37	,05	7,77	,000

4. Por último, la varianza de las pendientes (0,37), aunque se ha reducido prácticamente a la mitad, sigue siendo significativamente distinta de cero (*sig.* < 0,0005). Por tanto, una vez controlado el efecto de las variables *tto* y *cbasal*, todavía permanecen las diferencias entre las pendientes de los diferentes pacientes.

Apéndice 4

El tamaño muestral en los modelos multinivel

Al igual que ocurre con otros modelos lineales, tanto las estimaciones de los parámetros de un modelo multinivel como los contrastes que se aplican para valorar la significación estadística de esas estimaciones se realizan tomando como referencia un tamaño muestral concreto. Pero identificar el tamaño muestral *efectivo* en el que se basan esas estimaciones y contrastes es más complicado en un modelo multinivel que en otro tipo de modelos. Esto se debe, en parte, a que *cada nivel tiene su propio tamaño muestral*, pero, sobre todo, a que *la dependencia existente entre las observaciones anidadas produce pérdida de información*.

Prestar atención al tamaño muestral es importante por dos razones. En primer lugar, porque conviene conocer el tamaño muestral necesario para poder aplicar un modelo multinivel, es decir, el tamaño muestral necesario para que las estimaciones sean insesgadas y para que no haya problemas de convergencia. En segundo lugar, porque también conviene conocer el tamaño muestral con el que se consigue minimizar la probabilidad de cometer errores Tipo I y II al calcular la significación estadística de las estimaciones.

Antes de entrar en los detalles conviene comenzar señalando una cuestión de tipo general: cada efecto se valora tomando como referencia principal el tamaño muestral del nivel al que pertenece. En un estudio multinivel con 2.000 pacientes procedentes de 40 hospitales (en promedio, 50 pacientes por hospital), para contrastar el efecto de una variable del nivel 1 el tamaño muestral de referencia es el del nivel 1 (2.000 pacientes) y para contrastar el efecto de una variable del nivel 2 el tamaño muestral de referencia es el del nivel 2 (40 hospitales). Pero la presencia de efectos cruzados (interacciones entre variables de distinto nivel) y el hecho de tener que estimar simultáneamente parámetros de efectos fijos y parámetros de efectos aleatorios no contribuye precisamente a simplificar las cosas.

Convergencia

Si se utilizan en torno a 50 grupos y al menos 5 casos por grupo no suele haber problemas de convergencia. Conforme el modelo se va complicando por la incorporación de nuevos efectos aleatorios es posible que sean necesarios más casos para eliminar por completo los problemas de convergencia (ver Raudenbush, 2008).

Precisión de las estimaciones

Aunque en todo lo relativo al tamaño muestral en el ámbito de los modelos multinivel falta mucho por investigar, los resultados de los estudios de simulación disponibles ya permiten hacer

algunas recomendaciones. Quizá la más citada de estas recomendaciones sea la de Maas y Hox (2004), quienes sugieren utilizar al menos 20 grupos y al menos 30 casos por grupo. No obstante, Hox (2010), tras una completa revisión de los estudios disponibles, recomienda utilizar al menos 30 grupos y al menos 30 casos por grupo. No obstante, el número de casos por grupo parece no afectar de forma importante a las estimaciones cuando se tienen suficientes unidades del segundo nivel (Bell, Ferron y Kromrey, 2008; Bell, Morgan, Kromrey y Ferron, 2010).

Estas recomendaciones de tipo general necesitan ser matizadas. Existe bastante acuerdo en que las estimaciones de los parámetros de efectos fijos y de sus errores típicos suelen ser insesgadas sin necesidad de que el tamaño muestral sea grande en ninguno de los niveles (Bell, Morgan, Schoeneberger y Loudermilk, 2010). Los problemas surgen al estimar los parámetros de efectos aleatorios y, en particular, sus errores típicos. En un estudio de simulación con 30, 50 y 100 grupos (tamaños muestrales del nivel 2), y 5, 30 y 50 casos por grupo (tamaños muestrales del nivel 1), Maas y Hox (2005) concluyen que, aunque los coeficientes de regresión (parámetros de efectos fijos) y sus errores típicos se han estimado sin sesgo en todas las condiciones simuladas, los errores típicos de las varianzas del nivel 2 (la varianza de las medias y la varianza de las pendientes) son infra-estimados cuando el número de grupos del nivel 2 es menor de 100. La consecuencia de infra-estimar estos errores típicos es que aumenta la probabilidad de cometer errores Tipo I (con 30 grupos, por ejemplo, la infra-estimación es del 15%, lo cual conduce a una tasa de errores Tipo I del 8,9%).

En general, aunque 30 grupos y 5 casos por grupo puede ser suficiente para obtener estimaciones insesgadas de los parámetros de efectos fijos, estimar correctamente sus errores típicos requiere utilizar en torno a 50 grupos. Y para estimar correctamente los parámetros de efectos aleatorios y sus errores típicos es recomendable utilizar en torno a 100 grupos.

Potencia estadística

El tamaño muestral necesario para alcanzar una potencia aceptable depende del tipo de efecto evaluado: los efectos fijos requieren menos casos que los efectos aleatorios; y los efectos individuales menos casos que las interacciones entre variables de distinto nivel.

La potencia observada no suele alcanzar el nivel deseado de 0,80 cuando el número de grupos y el número de casos por grupo es muy pequeño. Bell, Morgan, Schoeneberger y Loudermilk (2010) señalan que, al evaluar los efectos fijos, únicamente se alcanza una potencia de 0,80 o mayor con 30 grupos y 20-40 casos por grupo; con tamaños muestrales más pequeños solo se alcanza una potencia aceptable si se trabaja con efectos de tamaño muy grande; y, en lo relativo a los efectos aleatorios, concluyen que no es nada fácil alcanzar una potencia aceptable. Hox (2010) sugiere que, en el caso de los efectos fijos, puede alcanzarse una potencia aceptable con 50 grupos y 5 casos por grupo; en el caso de los efectos aleatorios es necesario utilizar entre 100 y 200 grupos y 10 casos por grupo.

Por supuesto, la mejor forma de concretar el tamaño muestral que debe utilizarse en cada estudio concreto para alcanzar la potencia deseada consiste en realizar los cálculos pertinentes. Hox (2010) y Scherbaum y Ferreter (2009) explican cómo hacer esto. Y los programas informáticos *PinT* y *OptimalDesign*, ambos gratuitos, permiten hacer estos cálculos con suma facilidad (si bien Twisk, 2006, recomienda realizar los cálculos relativos a la potencia con mucha cautela). El programa *PinT* está diseñado para modelos de dos niveles; puede descargarse de la siguiente dirección: “<http://www.stats.ox.ac.uk/~snijders/multilevel.htm>”; se basa en el trabajo de Snijders y Bosker (1993). El programa *Optimal-Design* (Raudenbush, Spybrook, Congdon, Liu y Martínez, 2011) es, quizá el más extendido y también el más flexible y fácil de utilizar; puede descargarse de “<http://www.wtgrantfdn.org/resources/consultation-service-and-optimal-design>”.

Efecto del diseño

La dependencia entre las observaciones de una estructura multinivel produce, en mayor o menor medida, lo que se ha dado en llamar un **efecto del diseño** (*ED*). Este efecto refleja cómo son las cosas cuando las observaciones están anidadas en comparación con cómo son cuando no lo están. Puesto que un conjunto de observaciones dependientes contienen menos información que el mismo conjunto de observaciones independientes, la principal consecuencia del efecto del diseño es que el tamaño muestral *efectivo* (el tamaño muestral que debería utilizarse para fijar los grados de libertad, para calcular los errores típicos y para obtener la significación estadística de las estimaciones) suele ser distinto (generalmente más pequeño) del tamaño muestral *nominal*.

La pérdida de información derivada del hecho de trabajar con observaciones anidadas depende del modelo concreto que se está ajustando. Para varios modelos, incluido el modelo nulo o incondicional (ver [4.10]), el efecto del diseño puede estimarse mediante

$$ED = 1 + (n - 1)C_{Ci} \quad [4.30]$$

donde C_{Ci} es el coeficiente de correlación intraclase y n es el tamaño muestral medio de los grupos. El resultado de la ecuación [4.30] es tanto mayor cuanto mayor es la pérdida de información debida a la relación existente entre las observaciones, es decir, cuanto mayor es el valor del C_{Ci} . Un valor $ED = 1$ indica que no existe pérdida de información ($C_{Ci} = 0$). En nuestro ejemplo sobre 379 pacientes sometidos a tratamiento antidepresivo, al aplicar el modelo *nulo* (el modelo que únicamente incluye el efecto de los centros) hemos obtenido para el C_{Ci} un valor de 0,34 (ver los resultados del apartado *Análisis de varianza: un factor de efectos aleatorios*). Y el tamaño muestral medio de cada centro es $379/11 = 34,45$. Aplicando [4.30] obtenemos

$$ED = 1 + (34,45 - 1)(0,34) = 12,37$$

Este valor es una cuantificación de la pérdida de información que se produce en el primer nivel de nuestro diseño (modelo nulo) por el hecho de estar trabajando con pacientes agrupados en centros. Puesto que se trata de un valor mayor que 1, sabemos que se está perdiendo información, pero este valor por sí solo no permite precisar la magnitud de esa pérdida. No obstante, el valor del efecto del diseño puede utilizarse para corregir el tamaño muestral *nominal*, es decir, para obtener el tamaño muestral *efectivo*:

$$N_{efectivo} = N / ED \quad [4.31]$$

En nuestro ejemplo, el tamaño muestral *efectivo* ($379/12,37 = 30,64 \approx 31$) es sensiblemente más pequeño que el tamaño muestral *nominal* (379). En principio, este es el tamaño muestral que debería utilizar un análisis de regresión clásico para no infra-estimar los errores típicos de los coeficientes de regresión y para calcular correctamente la significación de los mismos. Pasar de 379 a 31 casos puede parecer una penalización exagerada solo por el hecho de estar trabajando con observaciones anidadas, pero esto es justamente lo que cabe esperar cuando se tienen pocos grupos (centros) y, comparativamente, muchos casos (pacientes) dentro de cada grupo. Para un mismo tamaño muestral, la penalización es tanto menor cuanto mayor es el número de grupos. Y cuanto mayor es el número de casos por grupo, mayor es también la pérdida de información y, consecuentemente, la penalización que hay que aplicar. Por lo general, es más informativo trabajar con 1.000 pacientes repartidos en 100 hospitales que con los mismos 1.000 pacientes repartidos en 20 hospitales (ver Snijders y Bosker, 1999).

Todo lo anterior se refiere al primer nivel (el nivel de los pacientes). En relación con el segundo nivel, el tamaño muestral viene dado por el número de grupos definidos por la variable

contextual (los centros). Con 11 centros, 11 es el tamaño muestral que sirve de referente para estimar los coeficientes y parámetros de covarianza del segundo nivel y para valorar la significación estadística de los mismos. Y la potencia en ese nivel solo mejora aumentando el número de unidades de ese nivel.

Por último, cuando se utiliza un modelo multinivel para analizar los datos de un diseño de medidas repetidas (donde las medidas repetidas son las unidades del primer nivel y los sujetos son las unidades del segundo nivel; recordemos que a estos modelos se les suele llamar *curvas de crecimiento*), lo habitual es que las unidades del primer nivel sean poco numerosas; es posible, incluso, que en un diseño de estas características solamente haya dos medidas, como en un diseño antes-después. Y con menos de 5 medidas (unidades del primer nivel) puede haber problemas de convergencia, falta de potencia y aumento de la tasa de errores Tipo I al contrastar la significación de los efectos aleatorios. Raudenbush (2008) señala que la potencia de este tipo de estudios depende, no sólo del número de medidas repetidas, sino del distanciamiento entre ellas y del tamaño del coeficiente de correlación intraclase.

5

Regresión logística (I). Respuestas dicotómicas

El **análisis de regresión logística** sirve para pronosticar una variable dependiente categórica a partir de una o más variables independientes de cualquier tipo (categóricas o cuantitativas).

La variable dependiente de una regresión logística puede ser dicotómica (regresión **binaria**) o politómica (regresiones **nominal** y **ordinal**). En este capítulo nos centraremos en la regresión logística *binaria*; en el próximo nos ocuparemos de las regresiones *nominal* y *ordinal*.

En un análisis de regresión logística *binaria* se tiene, en primer lugar, una variable dicotómica que define dos grupos: los pacientes que se recuperan y los que no, los clientes que devuelven un crédito y los que no, los ciudadanos que votan y los que no, etc.; esta variable dicotómica es la *variable dependiente* o *respuesta*, es decir, la variable cuyos valores se desea pronosticar. Y para efectuar esos pronósticos se tiene, en segundo lugar, una o más variables en las cuales se supone que se diferencian los grupos definidos por la variable dicotómica; estas variables en las que se supone que se diferencian los grupos son las *variables independientes* o *covariables* del análisis.

Al igual que el análisis de regresión lineal, el de regresión logística permite obtener una serie de pesos o coeficientes que informan sobre la contribución individual de cada variable independiente a la diferenciación entre los grupos y que permiten obtener pronósticos (en forma de probabilidades) que sirven para clasificar a los sujetos¹.

¹ El *análisis de regresión logística* comparte con el *análisis discriminante* el objetivo de generar pronósticos para clasificar a los sujetos en grupos. Pero el análisis de regresión logística se basa en supuestos menos exigentes que el análisis discriminante.

La utilidad de un análisis de estas características radica, desde luego, en la posibilidad de trabajar con respuestas dicotómicas (omnipresentes en el ámbito de las ciencias sociales y de la salud). Pero, además, los modelos diseñados para trabajar con respuestas dicotómicas constituyen la base de otros modelos más complejos. Los modelos que se utilizan para analizar respuestas nominales y ordinales se basan en la estimación simultánea de varios modelos para respuestas dicotómicas. Y cuando la variable dependiente es una frecuencia se utiliza una mezcla de modelos entre los cuales los diseñados para respuestas dicotómicas desempeñan un importante rol.

En este capítulo estudiaremos cómo ajustar un modelo de regresión logística, cómo valorar la calidad del ajuste, cómo interpretar los coeficientes del modelo, cómo realizar pronósticos y cómo chequear los supuestos del análisis. Para ampliar estos aspectos y otros que no trataremos aquí pueden consultarse los excelentes trabajos de Hosmer y Lemeshow (2000), Kleinbaum y Klein (2002), Kutner, Nachtsheim, Neter y Li (2005), Menard (2001) y Pampel (2000).

Regresión con respuestas dicotómicas

Aunque la regresión logística permite modelar respuestas nominales con cualquier número de categorías, comenzaremos con el caso más simple, es decir, con el modelo para una respuesta dicotómica, el cual no solo es el más utilizado en la práctica, sino que sirve de base para los demás.

Cuando se trabaja con variables dicotómicas (acierto-error, recuperados-no recuperados, a favor-en contra, comprar-no comprar, presencia-ausencia, etc.) es completamente irrelevante utilizar unos u otros códigos para identificar las categorías de la variable (es solo una cuestión de conveniencia), pero lo habitual es utilizar el código 1 para el acierto, los recuperados, la presencia, etc., y el código 0 para el error, los no recuperados, la ausencia, etc. En una variable de estas características, la probabilidad de cualquiera de sus dos valores es complementaria de la probabilidad del otro. Es decir, siendo Y una variable dicotómica con valores 0 y 1, se verifica

$$P(Y=1) = 1 - P(Y=0) \quad [5.1]$$

Por tanto, saber lo que ocurre con una cualquiera de las dos categorías implica saber lo que ocurre con la otra. Centrémonos en la categoría 1 y hagamos

$$E(Y) = P(Y=1) = \pi_1 \quad [5.2]$$

Esto significa que en una variable dicotómica codificada con “unos” y “ceros”, la media o valor esperado de la variable es la proporción de “unos”. Pero también significa que, a diferencia de lo que suele hacerse con una respuesta cuantitativa, con una respuesta dicotómica no interesa describir o pronosticar los valores concretos de la variable (los cuales sabemos que son intrínsecamente irrelevantes), sino la probabilidad de pertenecer a una de las dos categorías de la variable. Ahora bien, para explicar o pronosticar esta probabilidad pueden utilizarse diferentes estrategias. Veamos.

La función lineal

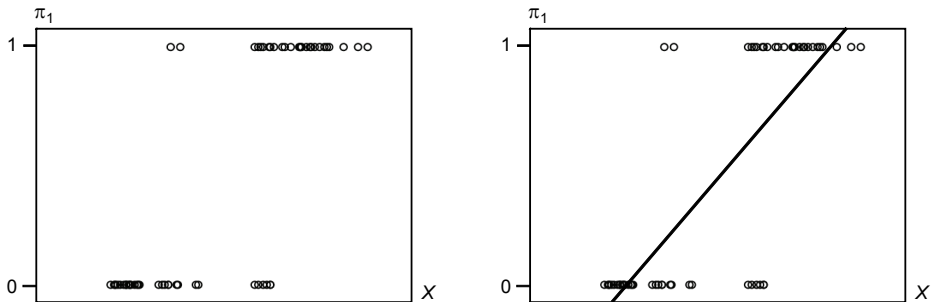
Una posible forma de modelar una respuesta dicotómica consiste en asumir que π_1 está linealmente relacionada con X y aplicar el modelo clásico de regresión lineal:

$$\pi_1 = \beta_0 + \beta_1 X \quad [5.3]$$

(en caso necesario, revisar el Capítulo 10 del segundo volumen). Los pronósticos que ofrece la ecuación [5.3] para π_1 forman una línea recta en el plano definido por las variables X e Y . El coeficiente β_0 es el punto en el que la recta corta el eje vertical; se le suele llamar *constante* o *intersección* (también, *ordenada en el origen*). El coeficiente β_1 define la *pendiente* de la recta, es decir, su inclinación respecto del eje horizontal; cuando no existe relación lineal, la recta es paralela al eje horizontal ($\beta_1 = 0$).

Aunque una ecuación lineal como la definida en [5.3] es muy útil para modelar una respuesta cuantitativa, no lo es tanto para modelar una respuesta dicotómica. Esto puede apreciarse fácilmente en los diagramas de dispersión de la Figura 5.1. El diagrama de la izquierda muestra los valores de una variable dicotómica Y respecto de una variable cuantitativa X cualquiera. Puesto que Y solo toma dos valores (0 y 1 en el ejemplo), los puntos del diagrama se encuentran alineados en dos filas. El diagrama de la derecha muestra la recta de regresión que ofrece la ecuación [5.3] para esta nube de puntos². Parece que, con variables dicotómicas, una línea recta no consigue hacer un buen seguimiento de la nube de puntos.

Figura 5.1. Diagrama de dispersión con recta de regresión lineal



Pero la calidad con la que una línea recta consigue resumir o representar una nube de puntos de estas características no es el único problema. El modelo de regresión lineal se basa en una serie de supuestos (linealidad, independencia, normalidad y homocedasticidad; ver Capítulo 10 del segundo volumen) que no se cumplen cuando la variable dependiente es dicotómica. En primer lugar, siendo Y una variable dicotómica, la relación

² En el eje vertical de este diagrama no están representados los dos valores de Y , sino la probabilidad de que cada caso tome el valor 1 o el valor 0; si los valores de X no se repiten, esas probabilidades siguen siendo 1 y 0 para cada caso, y el diagrama de dispersión es idéntico.

subyacente entre X e Y no puede ser lineal (ver Menard, 2001, págs. 7-11; los gráficos de la Figura 5.1 permiten apreciar esta circunstancia). En segundo lugar, los errores, es decir, las diferencias entre los valores de Y (0 y 1) y los pronósticos lineales π_1 que ofrece la ecuación [5.3] no son independientes de los valores de X : las puntuaciones bajas en X tienden a tener asociados errores negativos y las puntuaciones altas tienden a tener asociados errores positivos. En tercer lugar, las características de la variable dependiente hacen difícil que los errores puedan distribuirse normalmente, y esto afecta de forma importante tanto a los estadísticos que se utilizan para contrastar hipótesis sobre los coeficientes del modelo como a los intervalos de confianza que se construyen al estimar esos coeficientes. En cuarto lugar, la varianza de los errores no es constante para todo el rango de valores de X : la variabilidad de los errores es mayor cuando X toma valores intermedios que cuando toma valores extremos.

Además de estos problemas (algunos de los cuales podrían solucionarse utilizando muestras grandes y aplicando métodos de estimación alternativos a los mínimos cuadrados), ocurre que una recta de regresión lineal puede extenderse ilimitadamente por cualquiera de sus dos extremos conforme los valores de la variable independiente X van aumentando o disminuyendo. Consecuentemente, los pronósticos derivados de una ecuación lineal como la propuesta en [5.3] pueden tomar valores inaceptables (valores sin sentido). Esto es especialmente llamativo con respuestas dicotómicas. Puesto que la ecuación [5.3] está pronosticando probabilidades, todos los pronósticos deberían encontrarse en el rango 0-1. Sin embargo, para valores suficientemente extremos de X , la ecuación [5.3] puede ofrecer pronósticos imposibles, es decir, valores menores que 0 o mayores que 1. Por ejemplo, con los datos utilizados para obtener el diagrama representado en la Figura 5.1, la ecuación de regresión lineal ofrece pronósticos que oscilan entre $-0,098$ y $1,17$. Si Y es una variable dicotómica, π_1 no puede estar linealmente relacionada con un rango ilimitado de valores X .

La función logística

Las consideraciones del apartado anterior sugieren que una ecuación lineal no es una buena solución para modelar una respuesta dicotómica. Se obtienen mejores resultados con ecuaciones que, al definir una relación curvilínea entre X y π_1 , ofrecen pronósticos dentro del rango 0-1. Cualquier función de probabilidad acumulada monótona creciente cumple estos requisitos (relación curvilínea y pronósticos en el rango 0-1), pero la más utilizada para modelar respuestas dicotómicas es la **función logística**³, que para el caso de una sola variable independiente adopta la siguiente forma:

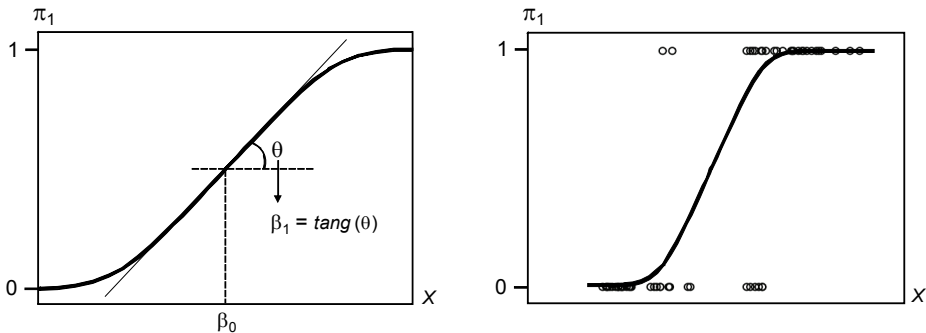
³ Otra función que suele recibir cierta atención en este contexto y que, con matices, ofrece resultados muy parecidos a la *logística* es la **función probit** (ver Apéndice 5). Esta función modela π_1 utilizando las probabilidades acumuladas asociadas a cada pronóstico lineal: $\pi_1 = F(\beta_0 + \beta_1 X)$, con la particularidad de que F se refiere a las probabilidades acumuladas de una distribución normal. La función se vuelve lineal cuando π_1 se multiplica por los valores inversos de F , es decir, $F^{-1}(\pi_1) = \beta_0 + \beta_1 X$. Esta función es menos flexible que la *logística* y no resulta nada fácil incluir en ella más de una variable independiente (ver Kutner y otros, 2005, págs. 559-560).

$$\pi_1 = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \quad [5.4]$$

(con $e = 2,71828$, base de los logaritmos naturales). Se trata de una función monótona con π_1 tendiendo a cero (si $\beta_1 < 0$) o a uno (si $\beta_1 > 0$) cuando X tiende a infinito. En la Figura 5.2 (izquierda) puede apreciarse la forma en “S” de esta función para $\beta_1 > 0$ (si $\beta_1 < 0$, la función sigue teniendo forma de “S”, pero invertida horizontalmente). Su utilidad para modelar probabilidades radica en el hecho de que, independientemente del valor que tome X , siempre ofrece valores comprendidos dentro del rango 0-1. Y, comparada con otras funciones, la logística es más versátil y ofrece resultados más fáciles de interpretar.

Al ajustar la función [5.4] al diagrama de dispersión representado en la Figura 5.1 se obtiene la curva que muestra la Figura 5.2 (derecha). El gráfico revela que la curva logística hace un seguimiento de la nube de puntos mejor que el que hace la recta de una ecuación lineal (Figura 5.1, derecha). Y no existen pronósticos imposibles: todos ellos se encuentran dentro del rango 0-1.

Figura 5.2. Curva de regresión logística (izquierda) con diagrama de dispersión (derecha)

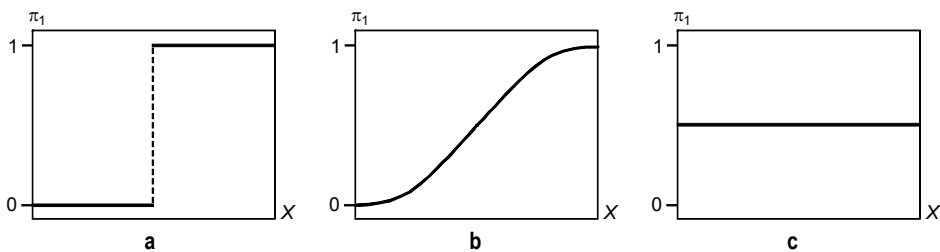


Ya sabemos que el ajuste de una recta a una nube de puntos va mejorando conforme se va alejando de cero el valor de su pendiente. Con una curva logística ocurre lo mismo. La Figura 5.3 muestra tres curvas logísticas ordenadas de forma decreciente por su capacidad para discriminar entre las dos categorías de la variable dicotómica Y . Cuando la variable independiente X es capaz de pronosticar correctamente la probabilidad de pertenecer a cada categoría de la variable dependiente Y , se obtiene una curva logística con mucha pendiente (es decir, un coeficiente β_1 alto en valor absoluto); cuando la variable independiente X no es capaz de pronosticar correctamente, se obtiene una curva sin pendiente o con muy poca pendiente (es decir, un coeficiente β_1 próximo a 0 en valor absoluto).

Una buena variable predictora (podríamos decir óptima) es aquella que permite obtener pronósticos (probabilidades) iguales o próximos a 1 para todos los casos en los

que se verifica $Y = 1$ y pronósticos iguales o próximos a 0 para todos los casos en los que se verifica $Y = 0$. La curva logística correspondiente a una variable de este tipo tiene forma de escalón (Figura 5.3.a). Por el contrario, una mala variable predictora (podríamos decir pésima) es aquella que pronostica a todos los sujetos el mismo o aproximadamente el mismo valor (la misma probabilidad), es decir, aquella que no contribuye en absoluto a distinguir entre las categorías de la variable dependiente. La curva correspondiente a una variable de este tipo tiene forma de línea paralela al eje de abscisas (Figura 5.3.c). Entre ambos extremos, es decir, entre la predicción óptima y la predicción pésima, existen múltiples curvas (la de la Figura 5.3.b es solo un ejemplo) que reflejan diferentes grados de precisión en la predicción y que se diferencian en el grado de inclinación, es decir, en el valor de β_1 .

Figura 5.3. Curvas logísticas ordenadas de mínima a máxima discriminación



La transformación logit

Con unas sencillas transformaciones se puede comprobar que, de la función logística propuesta en [5.4], se sigue

$$\text{odds}(Y=1) = \frac{\pi_1}{1 - \pi_1} = e^{\beta_0 + \beta_1 X} \quad [5.5]$$

Por tanto, la *odds* del suceso $Y=1$, es decir, el cociente entre π_1 y $1 - \pi_1$ (en caso necesario, revisar el concepto de *odds* en el Capítulo 3 del segundo volumen) permite simplificar la función logística propuesta en [5.4]. Y tomando el logaritmo de [5.5] se obtiene una ecuación lineal:

$$\log_e \left[\frac{\pi_1}{1 - \pi_1} \right] = \beta_0 + \beta_1 X \quad [5.6]$$

La parte izquierda de la ecuación [5.4] se conoce como *log-odds*, *transformación logit* o, simplemente, **logit**; y suele representarse abreviadamente mediante **logit($Y=1$)**. Por tanto,

$$\text{logit}(Y=1) = \beta_0 + \beta_1 X \quad [5.7]$$

Esta ecuación es la expresión habitual de lo que se conoce como **modelo de regresión logística** o, también, **modelo logit**⁴. Y puede extenderse fácilmente al caso de p variables independientes:

$$\text{logit}(Y=1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \quad [5.8]$$

Por tanto, en un modelo de regresión logística no se trabaja con los dos valores concretos de la variable dependiente Y (los cuales, tratándose de una variable dicotómica, son intrínsecamente irrelevantes), sino con la probabilidad de pertenecer a una de las dos categorías de la variable. Más concretamente, con el logaritmo de la *odds* de una de las dos categorías de la variable. El predictor lineal de un modelo de regresión logística (es decir, la parte derecha de la ecuación [5.8]) no pronostica $E(Y)$, sino el *logit* de $Y=1$. Es, por tanto, un modelo de la familia de los modelos lineales *generalizados* que utiliza una función de enlace *logit* (ver Apéndice 1). Y su utilidad radica precisamente en que permite expresar la transformación *logit* como una combinación *lineal* de efectos.

Es importante advertir que tanto $P(Y=1)$, como *odds* ($Y=1$), como *logit* ($Y=1$) están expresando la misma idea, pero en distinta escala. La correspondencia que muestra la Tabla 5.1 permite apreciar este hecho. Una *probabilidad* toma valores comprendidos entre cero y uno, y cada valor es simétrico de su complementario (a una probabilidad de 0,25 le corresponde un valor complementario de $1 - 0,25$). Una *odds* tiene un mínimo en cero y no tiene máximo (en teoría, $+\infty$); a una *probabilidad* de 0,50 le corresponde una *odds* de 1. Un *logit* no tiene ni mínimo ni máximo (en teoría, oscila entre $-\infty$ y $+\infty$); a una *probabilidad* de 0,50 le corresponde un *logit* de 0. Y, aunque una *probabilidad*

Tabla 5.1. Relación entre *probabilidad*, *odds* y *logit*

<i>Prob</i> ($Y=1$)	<i>Odds</i> ($Y=1$)	<i>Logit</i> ($Y=1$)
0,01	0,01	-4,60
0,10	0,11	-2,20
0,25	0,33	-1,10
0,50	1,00	0,00
0,75	3,00	1,10
0,90	9,00	2,20
0,99	99,00	4,60

⁴ Aunque ambas expresiones son equivalentes, cuando el modelo incluye alguna variable independiente cuantitativa suele utilizarse la expresión *modelo de regresión logística*; cuando todas las variables independientes son categóricas suele utilizarse la expresión *modelo logit*. Por tanto, cuando se ajusta un modelo de regresión *logística*, suele asumirse que los patrones de variabilidad se aproximan al número de casos; cuando se ajusta un modelo *logit*, suele asumirse que el número de casos es mayor que el de patrones de variabilidad. En el primer caso se habla de *datos no agrupados*; en el segundo, de *datos agrupados*. En el primer caso se asume que cada observación (que se considera única) sigue una distribución de Bernoulli, es decir, binomial con $n=1$ y $\pi=\pi_i$; en el segundo caso se asume que cada observación (cada patrón de variabilidad) sigue una distribución binomial con $n=n_h$ y $\pi=\pi_h$ (donde h se refiere a cada patrón de variabilidad, es decir, a cada combinación distinta entre las categorías de las variables independientes).

tiene una interpretación más fácil e intuitiva que una *odds*, y ésta más fácil e intuitiva que un *logit*, la transformación *logit* permite aprovechar las ventajas de trabajar con un modelo lineal.

Regresión logística binaria o dicotómica

En este apartado veremos cómo ajustar e interpretar un modelo de regresión logística con una variable dependiente dicotómica (ecuaciones [5.7] y [5.8]). En el próximo capítulo veremos cómo hacerlo con una variable dependiente politómica.

Al igual que en cualquier otro modelo de regresión, la selección de las variables independientes que formarán parte de un modelo de regresión logística puede hacerse a partir de criterios teóricos (en cuyo caso suele aplicarse una estrategia de inclusión forzosa de variables) o a partir de criterios estadísticos (en cuyo caso suele aplicarse algún método de selección por pasos). Veremos cómo hacer ambas cosas. Pero, cualquiera que sea la estrategia por la que se opte, una vez elegidas las variables, cubrir los diferentes objetivos del análisis requiere abordar tres tareas básicas: (1) valorar el ajuste global (es decir, valorar si las covariables incluidas en el modelo, tomadas juntas, están o no significativamente relacionadas con la variable dependiente) y estimar la fuerza o magnitud de la relación; (2) contrastar la significación individual de los coeficientes de regresión para identificar qué variables contribuyen al ajuste del modelo y en qué medida lo hace cada una; y (3) estudiar la adecuación del modelo chequeando los supuestos en los que se basa e indagando si existen casos atípicos e influyentes. Nos centraremos primero en las dos primeras tareas y dejaremos el chequeo de los supuestos para más tarde.

Para realizar estas tareas puede recurrirse a diferentes procedimientos SPSS: **Regresión logística binaria**, **Regresión logística multinomial**, **Regresión ordinal** y **Modelos lineales generalizados**. Nos centraremos principalmente en el primero de ellos, que es el que ha sido específicamente diseñado para el análisis de respuestas dicotómicas y el que ofrece la información más completa.

Los ejemplos que se proponen en este capítulo se basan en el archivo *Tratamiento adicción alcohol*, el cual puede descargarse de la página web del manual. El archivo contiene datos de 84 pacientes con problemas de alcoholismo que han participado en un programa de desintoxicación. Vamos a utilizar estos datos para averiguar si hay alguna variable que ayude a explicar o pronosticar la recuperación de los pacientes.

La variable que identifica a los pacientes recuperados es *recuperación*, una variable dicotómica⁵ con códigos 0 = “no” y 1 = “sí” (se han clasificado como recuperados los pacientes que no han recaído en los 18 meses siguientes a la finalización del trata-

⁵ Si se utiliza una variable dependiente politómica (más de dos categorías) con el procedimiento **Regresión logística binaria**, el SPSS emite una advertencia indicando que la variable seleccionada tiene más de dos categorías y que no es posible llevar a cabo el análisis. Para poder utilizar este procedimiento cuando la variable dependiente tiene más de dos categorías es necesario filtrar previamente los casos que pertenecen a las dos categorías con las que se desea trabajar o, alternativamente, recodificar la variable original haciéndole tomar solo dos valores, cuando esto tenga sentido.

miento). La categoría con el código más alto (1 en el caso de *recuperación*) desempeña un importante rol en el análisis. Los códigos asignados a las categorías de la variable dependiente no afectan al proceso de estimación (como es lógico, las estimaciones no pueden depender de los códigos que cada usuario decida utilizar); sin embargo, esos códigos condicionan por completo la interpretación de los resultados.

Para empezar a familiarizarnos con la variable *recuperación*, la Tabla 5.2 muestra su distribución de frecuencias. Los resultados indican que únicamente se han recuperado 36 de los 84 pacientes (el 42,9%).

Tabla 5.2. Distribución de frecuencias de la variable *recuperación*

	Frecuencia	Porcentaje	Porcentaje válido
Válidos No	48	57,1	57,1
Sí	36	42,9	42,9
Total	84	100,0	100,0

Una covariable (regresión simple)

Vamos a comenzar el estudio de la regresión logística con un modelo de regresión *simple*, es decir, con el modelo que incluye una sola covariable (a las variables independientes de la regresión logística se les suele llamar *covariables*). Y lo vamos a hacer con una covariable dicotómica para que se entienda mejor el significado de los coeficientes del modelo⁶. En concreto, vamos a comenzar con la variable *tto* (tratamiento). La mitad de los pacientes ha recibido un tratamiento *estándar* (a base de fármacos; código 0) y la otra mitad un tratamiento *combinado* (fármacos más psicoterapia; código 1).

Antes de comenzar el análisis vamos a averiguar si la variable *tto* está relacionada con la *recuperación*. La Tabla 5.3 muestra las frecuencias resultantes de cruzar ambas variables. Con el tratamiento estándar se recupera el 21,4% de los pacientes; con el combinado, el 64,3%.

Tabla 5.3. Frecuencias conjuntas de *tratamiento* por *recuperación*

			Recuperación		Total
			No	Sí	
Tratamiento	Estándar	Recuento	33	9	42
		% de Tratamiento	78,6%	21,4%	100,0%
	Combinado	Recuento	15	27	42
		% de Tratamiento	35,7%	64,3%	100,0%
Total		Recuento	48	36	84
		% de Tratamiento	57,1%	42,9%	100,0%

⁶ Por supuesto, para estudiar la relación entre dos variables dicotómicas no es necesario aplicar un modelo de regresión logística; estamos adoptando esta circunstancia como punto de partida porque creemos que de esta forma es más fácil entender los detalles del análisis.

Al contrastar la hipótesis de independencia mediante el estadístico X^2 de Pearson se obtiene un nivel crítico $p < 0,0005$ que delata una relación significativa entre ambas variables. Y la *odds ratio*, es decir, el cociente entre la *odds* de recuperarse con el tratamiento combinado ($27/15 = 1,800$) y la *odds* de recuperarse con el tratamiento estándar ($9/33 = 0,273$), vale $1,800/0,273 = 6,60$. Veremos que este valor desempeña un papel central en la interpretación de los resultados de la regresión logística.

Veamos cómo ajustar con el SPSS un modelo de regresión logística para pronosticar la *recuperación* de los pacientes a partir del tratamiento recibido (*tto*):

- Seleccionar la opción **Regresión > Logística binaria** del menú **Analizar** para acceder al cuadro de diálogo *Regresión logística*.
- Trasladar la variable *recuperación* al cuadro **Dependiente** y la variable *tto* a la lista **Covariables** (aunque el SPSS no establece restricciones en el tipo de *covariables* que pueden incluirse en el análisis, la variable *dependiente* debe ser *dicotómica*).

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestran las Tablas 5.4 a 5.12. Aprovechando esta información, en los siguientes apartados se explica cómo valorar el ajuste del modelo, cómo contrastar la significación de los coeficientes y cómo interpretarlos.

Información preliminar

La primera tabla informa del número de casos válidos incluidos en el análisis y del número de casos excluidos por tener algún valor perdido, ya sea en la variable dependiente, en la covariable o en ambas (ver Tabla 5.4).

Tabla 5.4. Resumen de los casos procesados

Casos no ponderados ^a		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	84	100,0
	Casos perdidos	0	,0
	Total	84	100,0
Casos no seleccionados		0	,0
Total		84	100,0

a. Si está activada la ponderación, consulte la tabla de clasificación para ver el número total de casos.

La Tabla 5.5 muestra la codificación interna que utiliza el procedimiento para identificar las dos categorías de la variable dependiente: el procedimiento asigna el valor interno 0 a la categoría con el código menor y el valor interno 1 a la categoría con el código mayor. En nuestro ejemplo, los códigos asignados coinciden con los códigos originales de la variable *recuperación*. Esta codificación interna no afecta a las estimaciones de los coeficientes, ni a sus errores típicos ni a su significación, pero es imprescindible conocerla para poder interpretar correctamente los resultados.

Tabla 5.5. Codificación de la variable dependiente

Valor original	Valor interno
No	0
Sí	1

Las Tablas 5.6 a 5.8 aparecen en el *Visor* bajo el título *Bloque 0 = Bloque inicial*. Estas tablas contienen información relativa al modelo *nulo*, es decir, al modelo que únicamente incluye el término constante. En las tablas de este bloque, una cabecera en la dimensión de las filas se encarga de recordar que se trata del paso 0. La información de este bloque o paso 0 no tiene utilidad en sí misma, sino que sirve de punto de referencia respecto del cual valorar cómo cambian las cosas cuando se van incorporando variables a la ecuación de regresión.

La Tabla 5.6 ofrece una clasificación de los casos en el paso 0. Esta tabla, conocida como *matriz de confusión*, recoge el resultado de cruzar los valores *observados* en la variable dependiente con los *pronosticados* por el modelo *nulo*. Puesto que el modelo *nulo* no incluye ninguna covariable, todos los casos son clasificados en la categoría más probable (la categoría a la que pertenecen más casos); en el ejemplo, la categoría de los *no recuperados*. De ahí que el porcentaje de casos correctamente clasificados (57,1%) coincida con el porcentaje de casos que pertenecen a esa categoría.

Tabla 5.6. Resultados de la clasificación en el paso 0 (matriz de confusión)

Observado		Pronosticado		
		Recuperación		% correcto
		No	Sí	
Paso 0	Recuperación No	48	0	100,0
	Sí	36	0	,0
Porcentaje global				57,1

La Tabla 5.7 ofrece una estimación de la *constante* del modelo ($-0,29$) junto con varios estadísticos asociados a esa estimación. La tabla también incluye el nivel crítico (*sig.*) resultante de contrastar la hipótesis nula de que el valor poblacional de la constante es cero. De momento (estamos en el paso 0), la constante es el único término presente en el modelo: $\text{logit}(\text{recuperarse} = 1) = \beta_0$. Y su valor se estima a partir de las frecuencias marginales de la variable dependiente:

$$\hat{\beta}_0 = \log_e(n_1/n_0) = \log_e(36/48) = -0,29$$

El valor negativo de $\hat{\beta}_0$ indica que la proporción de recuperados (la proporción de la categoría de referencia: $Y = 1$) es menor que la de no recuperados ($Y = 0$). Pero este valor está en escala logarítmica. Devolviéndolo a su escala natural se obtiene

$$\exp(\hat{\beta}_0) = e^{-0,29} = 0,75$$

Este valor se ofrece en la última columna de la tabla y no es otra cosa que la *odds* del suceso *recuperarse*, es decir, el cociente entre el número o proporción de recuperados y el número o proporción de no recuperados: $odds(recuperarse) = 36/48 = 0,75$. Y lo que indica esta *odds* es que el número o proporción de recuperados es un 75% del número o proporción de no recuperados (el resto de la información que contiene la tabla se explica más adelante; ver Tabla 5.12).

Tabla 5.7. Variables incluidas en la ecuación en el paso 0 (modelo nulo)

	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0 Constante	-,29	,22	1,70	1	,192	,75

La Tabla 5.8 informa de lo que ocurriría si se incorporaran al modelo cada una de las covariables elegidas. La tabla ofrece, para cada covariable, un contraste de la hipótesis de que su efecto es nulo (mediante el *estadístico de puntuación* de Rao, 1973). Puesto que, de momento, solo estamos utilizando la covariable *tto*, la tabla solo muestra información sobre esa covariable. Siguiendo la lógica habitual al contrastar hipótesis, si el nivel crítico asociado al estadístico de *puntuación* (*sig.*) es menor que 0,05, se puede rechazar la hipótesis nula (como en el ejemplo, pues $sig. < 0,0005$) y concluir que la correspondiente covariable contribuye significativamente a mejorar el ajuste del modelo nulo.

Tabla 5.8. Variables no incluidas en la ecuación en el paso 0

	Puntuación	gl	Sig.
Paso 0 Variables <i>tto</i>	15,75	1	,000
Estadísticos globales	15,75	1	,000

Ajuste global: significación estadística

Las Tablas 5.9 a 5.12 aparecen en el *Visor* bajo el título *Bloque 1: Método = Introducir* y contienen los resultados del modelo propuesto. El SPSS no ofrece la ecuación de regresión hasta el final (ver Tabla 5.12); en ese momento nos ocuparemos de ella.

Las Tablas 5.9 y 5.10 ofrecen la información necesaria para realizar una valoración global del modelo, es decir, para decidir si el conjunto de covariables incluidas en el análisis (de momento, solo la covariable *tto*) contribuyen o no a explicar una parte significativa de la variable dependiente (*recuperación*).

En regresión lineal esto se hace comparando sumas de cuadrados; en concreto, la suma de cuadrados de los residuos cuando el modelo incluye las variables independientes con esa misma suma de cuadrados cuando el modelo no incluye ninguna variable independiente (es decir, comparando la suma de cuadrados error o residual con la suma de cuadrados total). En regresión logística se hace algo parecido, pero, en lugar de utilizar sumas de cuadrados, se utilizan los logaritmos de las verosimilitudes.

Cuando las estimaciones se obtienen con el método de máxima verosimilitud, lo habitual es utilizar medidas de ajuste basadas en la **desvianza** ($-2LL$), la cual se obtiene a partir de las funciones de verosimilitud del modelo saturado y del modelo propuesto (ver, en el Capítulo 1, el apartado *Valorar la calidad o ajuste del modelo*). El ajuste de un modelo cualquiera es tanto mejor cuanto menor es su desvianza.

La desvianza del *modelo nulo* (la llamamos $-2LL_0$ por ser el valor de la desvianza en el paso 0) es equivalente a la suma de cuadrados total en un análisis de regresión lineal; representa, por tanto, el mayor desajuste posible. La desvianza del *modelo propuesto* (la llamamos $-2LL_1$ por ser el valor de la desvianza en el paso 1) es equivalente a la suma de cuadrados error en un análisis de regresión lineal. Y la diferencia entre ambas desvianzas es, en un análisis de regresión logística, un valor análogo a la suma de cuadrados debida a la regresión en un análisis de regresión lineal.

Esta diferencia, que expresa en qué medida el modelo propuesto consigue reducir el desajuste del modelo nulo, es la **razón de verosimilitudes** (G^2):

$$G_{0-1}^2 = -2LL_0 - (-2LL_1) \quad [5.9]$$

Conforme va aumentando el tamaño muestral, este estadístico se va aproximando a la distribución ji-cuadrado con los grados de libertad resultantes de restar el número de parámetros independientes⁷ de ambos modelos.

El SPSS ofrece el valor de $-2LL_1$ en la *tabla resumen del modelo* (ver Tabla 5.10) con el nombre *-2 log de la verosimilitud*, pero el valor de $-2LL_0$ no lo ofrece por defecto; para obtenerlo hay que marcar la opción **Historial de iteraciones** del subcuadro de diálogo *Regresión logística: Opciones*. En nuestro ejemplo, $-2LL_0 = 114,73$. La razón de verosimilitudes definida en [5.9] aparece en la tabla de *pruebas omnibus* (ver Tabla 5.9) con el nombre *chi-cuadrado*:

$$G_{0-1}^2 = 114,73 - 98,39 = 16,34$$

Este estadístico sirve para contrastar la hipótesis nula de que el modelo propuesto (el modelo que se está ajustando en el paso 1) no mejora el ajuste del modelo nulo (el modelo que se está ajustando en el paso 0). O, de forma equivalente, la hipótesis nula de que todos los coeficientes de regresión que incluye el modelo propuesto (excluida la constante) valen cero en la población:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0 \quad [5.10]$$

Con una sola covariable, la hipótesis [5.10] se reduce a $\beta_1 = 0$. Por tanto, la razón de verosimilitudes G^2 (*chi-cuadrado* en la Tabla 5.9) permite valorar si la covariable *tt* contribuye a mejorar el ajuste del modelo nulo. Puesto que el nivel crítico asociado a este estadístico (*sig.* < 0,0005) es menor que 0,05, se puede rechazar la hipótesis nula

⁷ El número de parámetros independientes de un modelo depende de la presencia de variables categóricas. Los modelos que solo incluyen covariables cuantitativas y dicotómicas tienen tantos parámetros como covariables más uno (el término constante). En los modelos que incluyen variables categóricas hay que añadir ($J - 1$) parámetros por cada variable categórica, siendo J el número de categorías de cada variable categórica.

$\beta_1 = 0$ y concluir que la covariable *tto* contribuye significativamente a mejorar el ajuste del modelo nulo⁸.

Los valores que ofrece la Tabla 5.9 (*paso, bloque y modelo*) permiten contrastar distintas hipótesis cuando se utiliza una estrategia secuencial de selección de variables (ver, más adelante, en el apartado *Regresión logística jerárquica o por pasos*).

Tabla 5.9. Pruebas *omnibus* sobre los coeficientes del modelo (contrastes de ajuste global)

	Chi-cuadrado	gl	Sig.
Paso 1 Paso	16,34	1	,000
Bloque	16,34	1	,000
Modelo	16,34	1	,000

Ajuste global: significación sustantiva

La Tabla 5.10 incluye, además de la desviación del modelo propuesto ($-2LL_1$), dos estadísticos R^2 que permiten valorar, no la significación estadística de las covariables incluidas en la ecuación, sino la fuerza o magnitud de la relación existente entre esas covariables y la variable dependiente (para una revisión de estas y otras medidas puede consultarse Menard, 2000).

En regresión lineal es habitual valorar la significación sustantiva de un modelo con el coeficiente de determinación $R^2_{XY} = SC_{\text{regresión}} / SC_{\text{total}}$. El coeficiente de determinación expresa, en escala de cero a uno, en qué medida el modelo de regresión consigue reducir los errores de predicción cuando, en lugar de pronosticar a todos los valores de Y su media, se utiliza la ecuación de regresión para realizar los pronósticos. Siguiendo con la analogía entre las sumas de cuadrados de la regresión lineal y los estadísticos $-2LL$ de la regresión logística, puede obtenerse una solución parecida al coeficiente de determinación (McFadden, 1974; Mennard, 2000; ver Long, 1997, para una revisión de varios estadísticos tipo R^2) mediante

$$R_L^2 = [-2LL_0 - (-2LL_1)] / (-2LL_0) = G_{0-1}^2 / (-2LL_0) \quad [5.11]$$

En el ejemplo, $R_L^2 = 16,34 / 114,73 = 0,14$. Este estadístico refleja la proporción de reducción de $-2LL_0$, es decir, la proporción en que el modelo propuesto (paso 1) consigue reducir la *desviación* o *desajuste* del modelo nulo (paso 0).

R_L^2 vale 0 cuando G_{0-1}^2 vale cero, es decir, cuando la reducción de la desviación es nula (lo cual significa que las variables incluidas en la ecuación no contribuyen en absoluto a reducir el desajuste) y se va aproximando a uno tanto más cuanto más se consigue reducir la desviación del modelo nulo.

⁸ En realidad, $-2LL$ no es una medida de *ajuste* sino de *desajuste* (pues el ajuste del modelo es tanto peor cuanto mayor es $-2LL$). Por tanto, la razón de verosimilitudes G^2 no está valorando en qué medida el modelo propuesto *mejora el ajuste* del modelo nulo, sino en qué medida el modelo propuesto *reduce el desajuste* del modelo nulo. Esto es algo parecido a lo que ocurre con el coeficiente de determinación en regresión lineal, el cual no indica en qué medida mejoran los pronósticos, sino en qué medida se reducen los errores de predicción.

El SPSS no incluye el estadístico R_L^2 , sino otros dos parecidos: Cox-Snell y Nagelkerke⁹. Ambos se parecen, conceptualmente, al coeficiente de determinación del análisis de regresión lineal, pero, dadas las características de la variable dependiente, debe tenerse muy presente que este tipo de estadísticos puede tomar valores bajos incluso cuando el modelo estimado pueda ser apropiado y útil. El estadístico de Nagelkerke indica que el modelo propuesto consigue reducir un 24% el desajuste del modelo nulo.

Tabla 5.10. Resumen del modelo (estadísticos de ajuste global)

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	98,39	,18	,24

Pronósticos y clasificación

Los estadísticos tipo R^2 del apartado anterior permiten valorar la calidad o ajuste de un modelo a partir de lo bien o mal que consigue pronosticar las probabilidades de cada categoría de la variable dependiente. Otra forma de valorar la calidad de un modelo consiste en comprobar cuántos casos consigue clasificar correctamente.

La clasificación de los casos se realiza a partir de las probabilidades pronosticadas. Y estas probabilidades se obtienen aplicando la ecuación propuesta en [5.4] tras sustituir los coeficientes β_0 y β_1 por sus correspondientes valores estimados $\hat{\beta}_0$ y $\hat{\beta}_1$, los cuales aparecen en la Tabla 5.12.

Veamos. La variable dependiente (Y) del ejemplo es *recuperación* (la categoría de referencia en el análisis es 1 = “sí”). La covariable (X) es *tto* y toma solo dos valores: 0 = “estándar” y 1 = “combinado”. Puesto que la covariable toma solo dos valores, la ecuación [5.4] solo genera *dos pronósticos distintos*. La probabilidad pronosticada $\hat{\pi}_1$ (es decir, la probabilidad de recuperación), es la probabilidad de recuperación cuando $X=0$ y cuando $X=1$:

$$\begin{aligned}\hat{\pi}_1 | (X=0) &= \frac{1}{1 + e^{-(-1,30 + 1,89 \times 0)}} = 0,21 \\ \hat{\pi}_1 | (X=1) &= \frac{1}{1 + e^{-(-1,30 + 1,89 \times 1)}} = 0,64\end{aligned}\quad [5.12]$$

La clasificación que recoge la Tabla 5.11 se basa en estas probabilidades. Las filas de la tabla clasifican los casos por su valor *observado* (el valor que toman en la variable *recuperación*); las columnas clasifican los casos por su valor *pronosticado* (la proba-

⁹ El estadístico de Cox y Snell (1989) se obtiene mediante $R_{\text{Cox-Snell}}^2 = 1 - (L_0/L_1)^{2/n}$, donde L_0 es la verosimilitud del modelo nulo (paso 0) y L_1 es la verosimilitud del modelo que se está ajustando (paso 1). El valor mínimo de este estadístico es cero (ajuste nulo), pero en caso de ajuste perfecto su valor máximo no es 1. Nagelkerke (1991) ha propuesto una modificación del estadístico de Cox y Snell que le permite alcanzar el valor 1 en caso de ajuste perfecto: $R_{\text{Nagelkerke}}^2 = R_{\text{Cox-Snell}}^2 / R_{\text{máx}}^2$, con $R_{\text{máx}}^2 = 1 - [L_0]^{2/n}$.

bilidad que les asigna la ecuación de regresión). Puesto que la probabilidad pronosticada (no olvidemos que se trata de la probabilidad asociada a la *recuperación*) es más alta con el tratamiento combinado (0,64) que con el estándar (0,21), los pacientes que han recibido el tratamiento combinado se han clasificado como *recuperados* y los pacientes que han recibido el tratamiento estándar se han clasificado como *no recuperados*¹⁰.

En la diagonal principal de la tabla se encuentran los casos que han resultado bien clasificados ($33 + 27 = 60$). Fuera de la diagonal principal se encuentran los casos que han resultado mal clasificados ($15 + 9 = 24$).

La última columna de la tabla informa del porcentaje de casos que han resultado correctamente clasificados en cada una de las dos categorías de la variable dependiente: *especificidad* = $(100)33/(33 + 15) = 68,8\%$; *sensibilidad* = $(100)27/(9 + 27) = 75,0\%$. La última fila de la tabla informa del *porcentaje total* de casos correctamente clasificados: $(100)60/(84) = 71,4\%$.

Los pacientes que se recuperan son algo mejor clasificados (*sensibilidad* = 75,0%) que los que no se recuperan (*especificidad* = 68,8%), pero como la clasificación se basa en dos pronósticos, no hay forma de cambiar esto. Cuando se trabaja con más de una covariable, el modelo genera muchos pronósticos distintos, particularmente si alguna de las covariables es cuantitativa. En estos casos, aunque mover el punto de corte no permite mejorar el porcentaje de casos correctamente clasificados, sí permite equilibrar la sensibilidad y la especificidad de la clasificación.

Tabla 5.11. Resultados de la clasificación en el paso 1 (matriz de confusión)

Observado			Pronosticado		
			Recuperación		% correcto
			No	Sí	
Paso 1 ^a	Recuperación	No	33	15	68,8
		Sí	9	27	75,0
	Porcentaje global				71,4

a. El punto de corte es ,50

Las frecuencias de una tabla de estas características pueden interpretarse aplicando alguna medida de asociación de las múltiples disponibles para analizar tablas de contingencias bidimensionales (ver el apartado *Medidas de asociación* del Capítulo 10 del primer volumen y, muy particularmente, el apartado *Medidas de asociación basadas en la reducción proporcional del error* del Apéndice 3 del segundo volumen). No obstante, debido a que cada una de estas medidas se centra en un aspecto diferente de la asociación, no parece estar del todo claro cuál de ellas ofrece una mejor solución (ver Menard, 2001, págs. 27-41).

¹⁰ Lógicamente, para efectuar esta clasificación es necesario establecer un punto de corte. La necesidad de establecer un punto de corte es más evidente cuando el modelo incluye varias covariables y a cada caso se le pronostica una probabilidad distinta. La clasificación se hace, por defecto, utilizando un punto de corte de 0,50 (se indica en una nota a pie de tabla), pero cualquier punto de corte comprendido entre 0,14 y 0,62, que son las dos probabilidades pronosticadas, habría llevado al mismo resultado.

Una forma sencilla, aunque no completamente libre de problemas, de aprovechar la información de una tabla de clasificación consiste en comparar los porcentajes de casos correctamente (o incorrectamente) clasificados que se obtienen con el modelo nulo (paso 0, Tabla 5.6) y con el modelo propuesto (paso 1, Tabla 5.11). En principio, cuanto mayor sea esta diferencia, más evidencia habrá de que las covariables incluidas en la ecuación de regresión contribuyen a mejorar el ajuste. En nuestro ejemplo, el porcentaje de casos correctamente clasificados es del 57,1% en el paso 0 y del 71,4% en el paso 1. Por tanto, al incorporar la información que aporta la covariable *tto*, el porcentaje de casos correctamente clasificados aumenta 14,3 puntos.

La significación estadística de ese aumento en el porcentaje de casos correctamente clasificados puede valorarse mediante

$$Z = (P_1 - P_0) / \sqrt{P_0(1 - P_0)/n} \quad [5.13]$$

(P_0 y P_1 se refieren a la proporción de casos correctamente clasificados en el paso 0 y en el paso 1, respectivamente). El estadístico Z se aproxima a $N(0, 1)$ conforme el tamaño muestral va aumentando y permite contrastar la hipótesis nula de que la proporción de casos correctamente clasificados en el paso 1 no difiere¹¹ de esa misma proporción en el paso 0. Podrá rechazarse esa hipótesis cuando Z sea mayor que el punto crítico de la distribución normal tipificada correspondiente a un nivel de confianza de 0,95 en un contraste unilateral derecho (es decir, cuando $Z > 1,64$). En nuestro ejemplo tenemos $P_0 = 0,571$, $P_1 = 0,714$ y $n = 84$ (ver Tablas 5.6 y 5.11). Por tanto,

$$Z = (0,714 - 0,571) / \sqrt{0,571(1 - 0,571)/84} = 2,65$$

Puesto que 2,65 es mayor que 1,64, puede concluirse que la proporción de casos correctamente clasificados es significativamente mayor en el paso 1 que en el paso 0.

Al interpretar el aumento en el porcentaje de casos correctamente clasificados debe tenerse en cuenta que un buen modelo desde el punto de vista de los pronósticos que ofrece (es decir, desde el punto de vista del tipo de ajuste del que informan los estadísticos tipo R^2) puede no ser un buen modelo desde el punto de vista de su capacidad para clasificar casos correctamente. Además, si la proporción de casos de una de las dos categorías de la variable dependiente es muy alta, el porcentaje de clasificación correcta será ya muy alto con el modelo nulo y no será nada fácil mejorarlo.

También debe tenerse en cuenta que una tabla de clasificación no contiene información acerca de cómo se distribuyen las probabilidades asignadas a cada grupo, es decir, no contiene información acerca de si las probabilidades individuales en las que se basa la clasificación están cerca o lejos del punto de corte. Y, obviamente, no es lo mismo clasificar a los sujetos a partir de probabilidades de recuperación de, por ejemplo, 0,95

¹¹ En realidad, el estadístico Z propuesto en [5.13] no es más que el estadístico que se utiliza en el contraste sobre una proporción (ver Capítulo 9 del primer volumen), con la particularidad de que, aquí, P_1 se interpreta como una variable que depende del modelo elegido (igual que la *proporción observada* en el contraste sobre una proporción) y P_0 como la proporción de referencia con la cual se compara P_1 (igual que la *proporción teórica* en el contraste sobre una proporción).

para los pacientes que han recibido el tratamiento combinado y 0,05 para los que han recibido el estándar, que clasificarlos con probabilidades de, por ejemplo, 0,55 y 0,45. En el primer caso hay cierta garantía de que los sujetos clasificados como recuperados se recuperarán y los clasificados como no recuperados no se recuperarán; en el segundo caso no existe tal garantía.

Por otro lado, el porcentaje de casos correctamente clasificados únicamente debe utilizarse como un criterio de ajuste cuando el objetivo del análisis sea clasificar a los sujetos. Si el objetivo del análisis es identificar las variables que contribuyen a entender el comportamiento de la variable dependiente, es preferible utilizar medidas de ajuste del tipo R^2 (ver Hosmer y Lemeshow, 2000, págs. 156-160).

Significación de los coeficientes de regresión

Recordemos que el modelo de regresión logística que estamos ajustando incluye la variable dependiente *recuperación* y la covariable *tto* (tratamiento):

$$\text{logit}(\text{recuperación} = 1) = \beta_0 + \beta_1(\text{tto})$$

La tabla de *variables incluidas en la ecuación* (Tabla 5.12) contiene las estimaciones de los coeficientes de regresión junto con la información necesaria para valorar su significación estadística e interpretarlos. La ecuación de regresión (es decir, la ecuación [5.7] tras estimar β_0 y β_1) queda de la siguiente manera:

$$\text{logit}(\text{recuperación} = 1) = -1,30 + 1,89(\text{tto})$$

El estadístico de Wald sirve para valorar la significación estadística de los coeficientes de regresión. Con variables cuantitativas y dicotómicas se obtiene elevando al cuadrado el cociente entre el valor del coeficiente (B) y su error típico ($E.T.$). Su distribución muestral se aproxima a ji-cuadrado con 1 grado de libertad. Este estadístico permite contrastar la hipótesis nula de que el coeficiente vale cero en la población:

$$H_0: \beta_j = 0 \quad [5.14]$$

Aplicando la estrategia habitual, si el nivel crítico (*sig.*) asociado al estadístico de Wald es menor que 0,05, se puede rechazar la hipótesis [5.14] y concluir que el valor poblacional del j -ésimo coeficiente de regresión es distinto de cero. El rechazo de esta hipótesis implica que la correspondiente covariable está significativamente relacionada con la variable dependiente.

Tabla 5.12. Variables incluidas en la ecuación en el paso 1 (modelo propuesto)

	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 ^a tto	1,89	,50	14,53	1	,000	6,60
Constante	-1,30	,38	11,94	1	,001	,27

a. Variable(s) incluida(s) en el paso 1: tto.

El estadístico de Wald es demasiado sensible al tamaño de los coeficientes (ver Hauck y Donner, 1977). Cuando el valor absoluto de un coeficiente es muy grande, también tiende a serlo su error típico. Y la consecuencia de esto es que el estadístico de Wald se vuelve conservador (tiende a rechazar la hipótesis nula [5.14] menos de lo que debería). En estos casos es preferible valorar la significación estadística de los coeficientes a partir del *cambio en la razón de verosimilitudes* (ver, más adelante, el apartado *Regresión logística por pasos*).

Interpretación de los coeficientes de regresión

Recordemos que el *modelo nulo* (paso 0; ver Tabla 5.7) y el *modelo propuesto*, es decir, el modelo que incluye la covariable *tto* (paso 1; ver Tabla 5.12) han quedado de la siguiente manera:

Modelo nulo (paso 0): $\text{logit}(\text{recuperación} = 1) = -0,29$

Modelo propuesto (paso 1): $\text{logit}(\text{recuperación} = 1) = -1,30 + 1,89(\text{tto})$

El valor de $\hat{\beta}_0$ cambia: pasa de $-0,29$ en el paso 0 a $-1,30$ en el paso 1. Y su valor exponencial pasa de $0,75$ en el paso 0 a $0,27$ en el paso 1. Su significado también cambia. En el modelo nulo (paso 0), $\exp(\hat{\beta}_0) = e^{-0,29} = 0,75$ es la *odds* de recuperarse: indica que el número total de recuperaciones (*recuperación* = 1) es un 75% del número total de no recuperaciones (*recuperación* = 0). En el modelo que incluye la covariable *tto* (paso 1), $\exp(\hat{\beta}_0) = e^{-1,30} = 0,27$ es la *odds* de recuperarse cuando todas las covariables (de momento, solo *tto*) valen cero. En la Tabla 5.3 puede comprobarse que, de los 42 pacientes que reciben el tratamiento estándar (*tto* = 0), solo se recuperan 9:

$$\text{odds}(\text{recuperación} | \text{estándar}) = 9/33 = 0,27$$

Este valor indica que, entre los pacientes que reciben el tratamiento estándar, el número de recuperaciones es un 27% del de no recuperaciones. O, de otra manera, entre los pacientes que reciben el tratamiento estándar, la recuperación se da un 73% menos de lo que se da la no recuperación.

El coeficiente $\hat{\beta}_1$, es decir, el coeficiente asociado a la covariable *tto*, vale 1,89. El valor de este coeficiente indica cómo cambia el *logit* de recuperarse (el pronóstico lineal de la ecuación logística) por cada unidad que aumenta *tto* (pasar del tratamiento estándar al combinado). El signo positivo del coeficiente indica que el *logit* de recuperarse aumenta cuando aumenta la covariable; por tanto, la probabilidad de recuperarse es mayor con el tratamiento combinado (*tto* = 1) que con el estándar (*tto* = 0).

La magnitud del coeficiente indica que el *logit* de recuperarse es 1,89 veces mayor con el tratamiento combinado que con el estándar. Pero razonar en escala *logit* es poco intuitivo. Devolviendo el valor del coeficiente a su escala natural (es decir, volviendo de [5.6] a [5.5]) se obtiene esa misma relación entre tratamientos, pero referida a las *odds*: $\exp(\hat{\beta}_1) = e^{1,89} = 6,60$ (ver última columna de la Tabla 5.12).

Así pues, la *odds* de recuperarse con el tratamiento estándar vale $e^{\hat{\beta}_0} = 0,27$; y la *odds* de recuperarse con el tratamiento combinado es 6,60 veces la de recuperarse con

el tratamiento estándar. Por tanto, 6,60 no es otra cosa que la *odds ratio* del suceso *recuperarse*, es decir, el cociente entre la *odds* de recuperarse con el tratamiento combinado y la *odds* de recuperarse con el tratamiento estándar (en caso necesario, revisar el concepto de *odds ratio* en el Capítulo 3 del segundo volumen). De otra forma, 6,60 es el valor por el que queda multiplicada la *odds* de recuperarse cuando se pasa del tratamiento estándar al combinado. Puesto que la *odds* de recuperarse con el tratamiento combinado es 6,60 veces la *odds* de recuperarse con el tratamiento estándar (un 560% mayor) y ésta vale 0,27 (utilizaremos 0,273 para evitar problemas de redondeo), la *odds* de recuperarse con el tratamiento combinado vale

$$odds(recuperación | combinado) = 6,60(0,27) = 1,80$$

Las *odds* obtenidas pueden utilizarse para interpretar los resultados en términos de **probabilidades**¹², lo cual suele ser más fácil de entender. Sabemos que existe una relación directa entre la probabilidad de un suceso y su *odds*. En concreto, $P = odds / (odds + 1)$:

$$P(recuperación | estándar) = 0,27 / (0,27 + 1) = 0,21$$

$$P(recuperación | combinado) = 1,80 / (1,80 + 1) = 0,64$$

(estas probabilidades coinciden con las ya pronosticadas por la ecuación logística en [5.12]). Así pues, la recuperación es más probable con el tratamiento combinado (0,64) que con el estándar (0,21). Pero no un 560% mayor, como ocurre con las correspondientes *odds*, sino un 205% (pues $100(0,64 - 0,21) / 0,21 = 204,7$). Por tanto, es muy importante no confundir la probabilidad de un suceso con su *odds*; ni un incremento en una con un incremento en la otra.

Más de una covariable (regresión múltiple)

Hasta ahora, por motivos didácticos, hemos explicado los aspectos básicos de la regresión logística ajustando un modelo con una sola covariable. No obstante, cuando se decide aplicar un modelo de regresión, lo habitual es intentar alcanzar el mayor ajuste posible incluyendo en él más de una covariable. En este apartado se explica cómo llevar a cabo un análisis de regresión múltiple (más de una covariable) con el SPSS.

¹² Para interpretar correctamente un coeficiente de regresión logística una vez devuelto a su métrica original hay que tener en cuenta que la *odds* de un suceso no es lo mismo que su *probabilidad*. Consecuentemente, la cantidad que aumenta la *odds* de un suceso no debe confundirse con la cantidad que aumenta su *probabilidad*. Veamos esto con algún ejemplo. Si la probabilidad de un suceso bajo la condición *A* vale 0,60, la *odds* de ese suceso vale $0,60/0,40 = 1,5$; si la probabilidad de ese suceso bajo la condición *B* vale 0,80, su *odds* vale $0,80/0,20 = 4$. Es decir, cuando la probabilidad de un suceso pasa de 0,60 a 0,80, su *odds* pasa de 1,5 a 4. Y la *odds ratio* expresa este aumento como un cambio proporcional: $4/1,5 = 2,67$, el cual indica que la *odds* del suceso ha aumentado un 167%. Es la *odds* del suceso la que aumenta un 167%, no su probabilidad, que aumenta un 33% (de 0,60 a 0,80). Otro ejemplo. Si la probabilidad de un suceso bajo la condición *A* vale 0,60, su *odds* vale $0,60/0,40 = 1,5$; si la probabilidad de ese suceso bajo la condición *B* vale 0,40, su *odds* vale $0,40/0,60 = 0,67$. Es decir, cuando la probabilidad de un suceso pasa de 0,60 a 0,40, su *odds* pasa de 1,5 a 0,67 (disminuye 0,83 puntos). La *odds ratio* expresa esta disminución como un cambio proporcional: $0,67/1,5 = 0,44$, el cual indica que la *odds* del suceso ha disminuido un 56%. Es la *odds* del suceso la que disminuye un 56%, no su probabilidad, que disminuye un 33% (de 0,60 a 0,40).

Seguimos con el mismo archivo (*Tratamiento adicción alcohol*) y la misma variable dependiente (*recuperación*) que en el primer ejemplo, pero con nuevas covariables:

- Seleccionar la opción **Regresión > Logística binaria** del menú **Analizar** para acceder al cuadro de diálogo *Regresión logística binaria*.
- Trasladar la variable *recuperación* al cuadro **Dependiente** y las variables *sexo*, *edad*, *años* (años consumiendo) y *tto* (tratamiento) a la lista **Covariables**.
- Pulsar el botón **Opciones** para acceder al subcuadro de diálogo *Regresión logística: Opciones* y marcar las opciones **Bondad de ajuste de Hosmer-Lemeshow** e **IC para exp(B)**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor* ofrece, entre otros, los resultados que muestran las Tablas 5.13 a 5.19.

Información preliminar

Toda la información que se obtiene en el paso 0 es idéntica a la obtenida en el apartado anterior (ver Tablas 5.4 a 5.7): el modelo nulo no cambia por elegir unas u otras covariables; siempre es el modelo que incluye únicamente el término constante. La información del paso 0 indica que el número total de casos válidos es 84 y que los códigos internos asignados a las categorías de la variable dependiente siguen siendo 1 para los pacientes que se recuperan y 0 para los no se recuperan. El único coeficiente que incluye el modelo nulo (la constante) vale $-0,29$, y su valor exponencial es $e^{-0,29} = 0,75$, el cual indica que el número total de recuperaciones es un 75% del número total de no recuperaciones. Además, la tabla de clasificación correspondiente al modelo nulo refleja un porcentaje de clasificación correcta del 57,1%.

Por último, todavía dentro del paso 0, se ofrece un avance de qué covariables tendrían un peso significativo de ser incluidas en el modelo (ver Tabla 5.13). El estadístico *puntuación* permite contrastar la hipótesis nula de que la correspondiente covariable no está relacionada con la variable dependiente. A las variables *sexo*, *años* y *tto* les corresponden niveles críticos menores que 0,05; por tanto, en principio, las tres variables son buenas candidatas para formar parte del modelo de regresión. Con la variable *edad* no ocurre lo mismo (*sig.* = 0,545). La última línea, *estadísticos globales*, permite contrastar la hipótesis de no relación entre la variable dependiente y las cuatro covariables tomadas juntas; el nivel crítico obtenido (*sig.* < 0,0005) permite rechazar esa hipótesis.

Tabla 5.13. Variables no incluidas en la ecuación en el paso 0

				Puntuación	gl	Sig.
Paso 0	Variables	sexo		6,68	1	,001
		edad		,37	1	,545
		años		12,69	1	,000
		tto		15,75	1	,000
	Estadísticos globales			29,42	4	,000

Ajuste global: significación estadística

La Tabla 5.14 ofrece una valoración del cambio que ha experimentado la *desvianza* del modelo nulo al incorporar las covariables *sexo*, *edad*, *años* y *tto*. A este cambio en la desvianza lo hemos llamado *razón de verosimilitudes* (G^2 ; ver ecuación [5.9]) y aparece en la Tabla 5.14 con el nombre *chi-cuadrado*. La razón de verosimilitudes permite contrastar la hipótesis nula de que todos los coeficientes de regresión (todos menos la constante) valen cero (ver hipótesis [5.10]). Por tanto, el estadístico *chi-cuadrado* que ofrece la Tabla 5.14 permite valorar si el modelo propuesto (el modelo en el paso 1) consigue reducir el desajuste del modelo nulo (el modelo en el paso 0).

Dado que el modelo se construye en un único paso, todas las entradas de la tabla (paso, bloque, modelo) están contrastando la misma hipótesis nula: que los coeficientes de regresión en que difieren el modelo 0 y el modelo 1 valen cero. En nuestro ejemplo, puesto que el nivel crítico asociado al estadístico *chi-cuadrado* (*sig.* < 0,0005) es menor que 0,05, se puede rechazar esa hipótesis nula y concluir que las variables elegidas, tomadas juntas, contribuyen a reducir el desajuste del modelo nulo.

Tabla 5.14. Pruebas *omnibus* sobre los coeficientes del modelo (contrastes de ajuste global)

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	34,63	4	,000
	Bloque	34,63	4	,000
	Modelo	34,63	4	,000

De las diferentes estrategias disponibles para valorar el ajuste de un modelo de regresión logística (ver Hosmer, Hosmer, Le Cessie y Lemeshow, 1997), el SPSS ofrece el estadístico de bondad de ajuste de Hosmer-Lemeshow (1980, 2000). En el ejemplo del apartado anterior no hemos solicitado este estadístico porque solo tiene sentido aplicarlo si el modelo que se está ajustando genera muchos pronósticos distintos, no unos pocos; y esto solo es posible si el modelo incluye muchas covariables y, particularmente, si alguna de ellas es cuantitativa.

La Tabla 5.15 contiene el estadístico de Hosmer-Lemeshow (*chi-cuadrado*) y su significación estadística; la Tabla 5.16 ofrece los datos a partir de los cuales se obtiene este estadístico. Aunque la forma concreta de calcular este estadístico admite algunas variantes, el SPSS lo hace dividiendo la muestra en 10 grupos del mismo tamaño a partir de sus probabilidades pronosticadas (el primer grupo lo forma el 10% de los casos con las probabilidades pronosticadas más bajas; el décimo grupo lo forma el 10% de los casos con las probabilidades pronosticadas más altas).

Tras esto se calculan dos tipos de frecuencias: las observadas y las esperadas (ver la Tabla 5.16). Las *frecuencias observadas* se obtienen contando el número de casos de cada grupo que pertenecen a cada categoría de la variable dependiente. Si $Y = 1$ (en nuestro ejemplo, *recuperación* = “sí”), las *frecuencias esperadas* se obtienen sumando las probabilidades pronosticadas $P(Y = 1)$ de todos los casos de cada grupo; si $Y = 0$ (en nuestro ejemplo, *recuperación* = “no”), la *frecuencia esperada* se obtiene sumando los

valores complementarios de las probabilidades pronosticadas $1 - P(Y = 1)$ de todos los casos de cada grupo. Se obtiene así una tabla de contingencias bidimensional de tamaño 10×2 (los 10 grupos y las dos categorías de la variable dependiente) con la particularidad de que cada casilla de la tabla contiene una frecuencia observada y su correspondiente frecuencia esperada.

Tabla 5.15. Prueba de *Hosmer-Lemeshow*

Paso	Chi-cuadrado	gl	Sig.
1	19,77	8	,011

Tabla 5.16. Tabla de contingencias para la prueba de *Hosmer-Lemeshow*

		Recuperación = No		Recuperación = Sí		Total
		Observado	Esperado	Observado	Esperado	
Paso 1	1	8	7,600	0	,400	8
	2	5	7,257	3	,743	8
	3	6	6,798	2	1,202	8
	4	7	6,237	1	1,763	8
	5	7	5,699	1	2,301	8
	6	6	4,812	2	3,188	8
	7	3	3,847	5	4,153	8
	8	6	2,913	2	5,087	8
	9	0	1,829	8	6,171	8
	10	0	1,007	12	10,993	12

Hosmer y Lemeshow han demostrado que puede utilizarse el estadístico X^2 de Pearson (que, en este caso, se aproxima a la distribución ji-cuadrado con 8 grados de libertad) para contrastar la hipótesis nula de que las frecuencias pronosticadas por el modelo se parecen a las observadas. En nuestro ejemplo, este estadístico toma el valor 19,77 y tiene asociado un nivel crítico (*sig.*) de 0,011. Por tanto, lo razonable es rechazar la hipótesis nula y concluir que el ajuste obtenido no es del todo satisfactorio. Sin embargo, debe tenerse en cuenta que la presencia de variables irrelevantes en la ecuación de regresión suele afectar de forma negativa a la precisión de esta prueba de ajuste. Según veremos enseguida, la variable *edad* no está contribuyendo significativamente al ajuste del modelo. Esto quiere decir que la variable *edad* podría ser excluida del análisis sin pérdida de ajuste. Y lo que ocurre al excluir del modelo la variable *edad* es que el estadístico *chi-cuadrado* de la Tabla 5.15 cambia de 19,77 a 14,04 y su nivel crítico de 0,011 a 0,081. Y, con estos nuevos resultados, la conclusión razonable es no rechazar la hipótesis de ajuste.

Para utilizar esta prueba de ajuste es necesario trabajar con muestras grandes y con covariables capaces de generar un pronóstico distinto para todos o casi todos los casos. Pero, al mismo tiempo, debe tenerse en cuenta que, puesto que el valor del estadístico *chi-cuadrado* es sensible al tamaño muestral, con muestras muy grandes podría llevar a rechazar la hipótesis de ajuste incluso con modelos que se ajustan bien a los datos.

Ajuste global: significación sustantiva

La Tabla 5.17 contiene varios estadísticos tipo R^2 que permiten valorar la intensidad de la relación entre la variable dependiente y el conjunto de covariables incluidas en el modelo (ver la ecuación [5.11] y la nota a pie de página número 9). Comparando estos resultados con los de la Tabla 5.10 se observa que tanto el estadístico de Cox y Snell como el de Nagelkerke toman valores sensiblemente más altos. Ahora, el estadístico de Nagelkerke vale 0,45 (21 centésimas más que cuando el modelo únicamente incluía la covariable *tto*): las covariables incluidas en el modelo consiguen reducir un 45% la desviación (el desajuste) del modelo nulo.

Tabla 5.17. Resumen del modelo (estadísticos de ajuste global)

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	80,10	,34	,45

La Tabla 5.18 muestra el resultado de la clasificación. Recordemos que esta clasificación se basa en las probabilidades pronosticadas (ver ecuación [5.12]). El modelo propuesto (paso 1), clasifica correctamente al 77,4% de los casos. Puesto que el modelo nulo (paso 0) clasificaba correctamente al 57,1% de los casos (ver Tabla 5.6), la clasificación basada en los pronósticos del modelo propuesto consigue mejorar el porcentaje de clasificación correcta en 20,3 puntos.

Al aplicar la ecuación [5.13] para valorar este aumento de 20,3 puntos porcentuales se obtiene $Z = 3,76$, y este valor que permite rechazar la hipótesis nula de que no ha habido cambio. Por tanto, puede concluirse que el modelo de regresión propuesto consigue clasificar correctamente un porcentaje de casos significativamente más alto que el modelo nulo.

Tabla 5.18. Resultados de la clasificación en el paso 1 (matriz de confusión)

Observado			Pronosticado		
			Recuperación		% correcto
			No	Sí	
Paso 1	Recuperación	No	40	8	83,3
		Sí	11	25	69,4
	Porcentaje global				77,4

Significación de los coeficientes de regresión

Recordemos que el modelo de regresión logística que estamos ajustando incluye la variable dependiente *recuperación* y cuatro covariables:

$$\text{logit}(\text{recuperación} = 1) = \beta_0 + \beta_1(\text{sexo}) + \beta_2(\text{edad}) + \beta_3(\text{años}) + \beta_4(\text{tto})$$

La tabla de *variables incluidas en la ecuación* (Tabla 5.19) contiene las estimaciones de los coeficientes de regresión (columna *B*) junto con la información necesaria para valorar su significación estadística e interpretarlos.

La significación de cada coeficiente se evalúa con el estadístico de *Wald*, el cual, recordemos, permite contrastar la hipótesis nula de que el correspondiente coeficiente vale cero en la población. En nuestro ejemplo, las covariables *sexo*, *años* y *tto* tienen asociados coeficientes de regresión significativamente distintos de cero (*sig.* < 0,05 en los tres casos), pero el coeficiente asociado a la variable *edad* no alcanza la significación estadística (*sig.* = 0,081). Por tanto, puede concluirse que todas las covariables, excepto la *edad*, contribuyen a reducir el desajuste del modelo nulo.

Tabla 5.19. Variables incluidas en la ecuación en el paso 1 (covariables: *sexo*, *edad*, *años* y *tto*)

		B	E.T.	Wald	gl	Sig.	Exp(B)	IC 95% para EXP(B)	
								Inferior	Superior
Paso 1	sexo	-1,63	,63	6,60	1	,010	,20	,06	,68
	edad	,10	,06	3,04	1	,081	1,10	,99	1,23
	años	-,29	,09	9,39	1	,002	,75	,62	,90
	tto	1,59	,57	7,67	1	,006	4,90	1,59	15,07
	Constante	,49	1,38	,13	1	,721	1,64		

Las covariables cuyos coeficientes de regresión no son significativamente distintos de cero conviene eliminarlas del modelo. Eliminar estas covariables no solo no altera la calidad del modelo (no empeora el ajuste) sino que ayuda a que las estimaciones del nuevo modelo sean más eficientes.

Al eliminar de nuestro modelo la variable *edad* (ver Tabla 5.20), el valor de los restantes coeficientes cambia ligeramente (recordemos que se trata de coeficientes de regresión *parciales*, es decir, de coeficientes cuyo valor estimado viene condicionado por el resto de coeficientes presentes en el modelo). El nuevo modelo de regresión queda de la siguiente manera :

$$\text{logit}(\text{recuperación} = 1) = 2,11 - 1,33(\text{sexo}) - 0,18(\text{años}) + 1,84(\text{tto}) \quad [5.15]$$

Y los correspondientes errores típicos son, todos ellos, ligeramente más pequeños. Lo cual viene a confirmar que, al eliminar de la ecuación una variable irrelevante, las estimaciones se vuelven más eficientes.

Tabla 5.20. Variables incluidas en la ecuación en el paso 1 (covariables: *sexo*, *años* y *tto*)

		B	E.T.	Wald	gl	Sig.	Exp(B)	IC 95% para EXP(B)	
								Inferior	Superior
Paso 1	sexo	-1,33	,59	5,06	1	,024	,26	,08	,84
	años	-,18	,07	7,34	1	,007	,84	,73	,95
	tto	1,84	,55	11,01	1	,001	6,27	2,12	18,54
	Constante	2,11	1,07	3,86	1	,049	8,23		

Interpretación de los coeficientes de regresión

El signo de los coeficientes de regresión refleja el sentido (positivo o negativo) de la relación entre cada covariable y la variable dependiente. Por tanto, la variable *tto* se relaciona positivamente con la *recuperación* (a más *tto* –pasar de 0 a 1–, más *recuperación*) y las variables *sexo* y *años* se relacionan negativamente con la *recuperación* (a más *sexo* –pasar de 0 a 1– y más *años consumiendo*, menos *recuperación*). Y, dado que estos coeficientes se encuentran en escala logarítmica, a los coeficientes positivos les corresponden valores exponenciales mayores que 1 y, a los negativos, valores exponenciales menores que 1. Veamos cómo se interpreta cada uno de estos coeficientes:

- **Coefficiente $\hat{\beta}_0$.** El valor de la constante (2,11) es el pronóstico, en escala *logit*, que ofrece el modelo de regresión cuando todas las covariables (en el ejemplo, *sexo*, *años* y *tto*) valen cero. Para que el valor exponencial de la constante ($e^{2,11} = 8,23$) tenga algún significado es imprescindible que el valor cero también tenga algún significado en todas las covariables. En nuestro ejemplo, esto ocurre con las variables *sexo* y *tto*, pero no hay sujetos con cero años de consumo.

Al recodificar la variable *años* asignando el valor cero a los pacientes con 14 años de consumo (el valor de la mediana), se obtiene para la constante $\hat{\beta}_0$ un valor de -0,42 (el resto de coeficientes no cambian). Este valor es el pronóstico, en escala *logit*, que la ecuación de regresión estima para las mujeres (*sexo* = 0) que han recibido el tratamiento estándar (*tto* = 0) y que llevan 14 años consumiendo (*años* = 0). La *odds* de recuperarse en estas pacientes vale $e^{-0,42} = 0,66$. Y esto significa que, en estas pacientes, el número de recuperaciones es un 66% del de no recuperaciones.

- **Coefficiente $\hat{\beta}_1$.** El signo negativo del coeficiente correspondiente a la variable *sexo* (-1,33) indica que la recuperación es más probable entre las mujeres (*sexo* = 0) que entre los hombres (*sexo* = 1). Y el valor exponencial del coeficiente ($e^{-1,33} = 0,26$) indica que la *odds* de recuperarse entre los hombres (*sexo* = 1) es un 26% de la *odds* de recuperarse entre las mujeres. De otra manera: la *odds* de recuperarse entre los hombres es un 74% menor que entre las mujeres.
- **Coefficiente $\hat{\beta}_2$.** El signo negativo del coeficiente estimado para la variable *años* (-0,18) significa que la probabilidad de recuperación disminuye cuando aumentan los años de consumo. El valor exponencial del coeficiente ($e^{-0,18} = 0,84$) indica que, con cada año más de consumo¹³, la *odds* de recuperarse disminuye un 16%.
- **Coefficiente $\hat{\beta}_3$.** Por último, el coeficiente estimado para la variable *tto* (1,84) indica que la recuperación aumenta cuando aumenta *tto*, es decir, cuando *tto* pasa de estándar (*tto* = 0) a combinado (*tto* = 1). El valor de la correspondiente *odds ratio* ($e^{1,83} = 6,27$) revela que la *odds* de recuperarse con el tratamiento combinado es

¹³ Con covariables cuantitativas como la variable *años* puede interesar interpretar la *odds ratio* asociada no a un valor (un año) sino a un intervalo de valores (un lustro, una década). En ese caso, la *odds ratio* asociada a un cambio de *k* unidades se obtiene mediante e^{kB} , siendo *B* el coeficiente de regresión estimado para el cambio de una unidad. En nuestro ejemplo, la *odds ratio* asociada a cinco años de consumo vale $e^{5(-0,18)} = 0,41$, lo cual indica que, por cada cinco años más de consumo, la *odds* de recuperarse disminuye un 59%.

6,27 veces la *odds* de recuperarse con el tratamiento estándar. De otra manera: una *odds ratio* de 6,27 indica que la *odds* de recuperarse con el tratamiento combinado es un 527% mayor que la de recuperarse con el tratamiento estándar.

Los **intervalos de confianza** que aparecen al final de la tabla indican entre qué valores se estima que se encuentran, con una confianza del 95%, los valores poblacionales de las *odds ratios* estimadas. Aunque estos intervalos no se refieren a los coeficientes de regresión sino a sus valores exponenciales, no se calculan a partir de éstos (que tienen una distribución muestral muy asimétrica), sino a partir de los coeficientes $\hat{\beta}_j$ (que se asume que se distribuyen normalmente). Para obtener estos intervalos de confianza, primero se calculan los límites correspondientes a β_j mediante:

$$IC_{\beta_j} = \hat{\beta}_j \pm |Z_{\alpha/2}| S_{\hat{\beta}_j} \quad [5.16]$$

Y, a continuación, se calculan los valores exponenciales de IC_{β_j} . Por ejemplo, para obtener, con $1 - \alpha = 0,95$, el intervalo de confianza correspondiente al coeficiente de regresión de la variable *años*, primero se aplica [5.16]:

$$IC_{\beta_{\text{años}}} = \hat{\beta}_{\text{años}} \pm |Z_{0,025}| S_{\hat{\beta}_{\text{años}}} = -0,18 \pm 1,96(0,07) = (-0,32; -0,04)$$

Estos límites se encuentran en escala logarítmica. Devolviéndolos a su escala natural se obtiene $e^{-0,32} = 0,73$ y $e^{-0,04} = 0,96$, valores que, salvo por detalles de redondeo, son justamente los que ofrece la Tabla 5.20 en las dos últimas columnas para los límites inferior y superior del intervalo de confianza al 95 %. Esto significa que, basándonos en el modelo de regresión propuesto, podemos estimar que la *odds* de recuperarse disminuye entre un 4% y un 27% por cada año más de consumo.

Pronósticos y clasificación

Los resultados de la clasificación ya los hemos presentado en la Tabla 5.18. Recordemos que esta tabla de clasificación se construye a partir de las probabilidades pronosticadas. Los pronósticos lineales se obtienen asignando valores a las covariables *sexo*, *años* y *tto* en la ecuación [5.15] (el modelo en el que se basa la clasificación de la Tabla 5.18 incluye también la covariable *edad*). El valor pronosticado más bajo corresponde a un hombre (*sexo* = 1) con el mayor número de años de consumo (*años* = 22) y que ha recibido el tratamiento estándar (*tto* = 0); el valor pronosticado más alto corresponde a una mujer (*sexo* = 0) con el menor número de años de consumo (*años* = 2) y que ha recibido el tratamiento combinado (*tto* = 1):

$$\begin{aligned} \text{logit}_{\text{más bajo}}(\text{recuperación} = 1) &= 2,11 - 1,33(1) - 0,18(22) + 1,84(0) = -3,18 \\ \text{logit}_{\text{más alto}}(\text{recuperación} = 1) &= 2,11 - 1,33(0) - 0,18(2) + 1,84(1) = 3,59 \end{aligned}$$

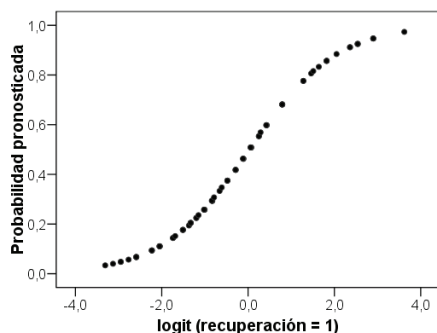
Estos pronósticos se encuentran en escala *logit*. Unas sencillas operaciones permiten transformarlos en *probabilidades* (más fáciles de interpretar). Al *logit* más bajo le corresponde la probabilidad más baja; al *logit* más alto, la probabilidad más alta:

$$\hat{\pi}_{1(\text{más baja})} = 1/(1 + e^{-(-3,18)}) = 0,04$$

$$\hat{\pi}_{1(\text{más alta})} = 1/(1 + e^{-(3,59)}) = 0,97$$

Por tanto, la probabilidad de recuperación estimada es muy baja (0,04) cuando se aplica el tratamiento estándar a hombres que llevan 22 años consumiendo y muy alta (0,97) cuando se aplica el tratamiento combinado a mujeres que llevan solo 2 años consumiendo. El resto de los pronósticos se obtienen de la misma manera que estos dos. En la Figura 5.4 están representadas las probabilidades que el modelo de regresión pronostica (eje vertical) para cada patrón de variabilidad (eje horizontal). La disposición de los puntos nos recuerda que estamos utilizando un modelo de regresión basado en la transformación *logit*. Los pronósticos 0,04 y 0,97 corresponden a los dos extremos de la nube de puntos.

Figura 5.4. Relación entre el *logit* de *Y* (pronóstico lineal) y las probabilidades pronosticadas



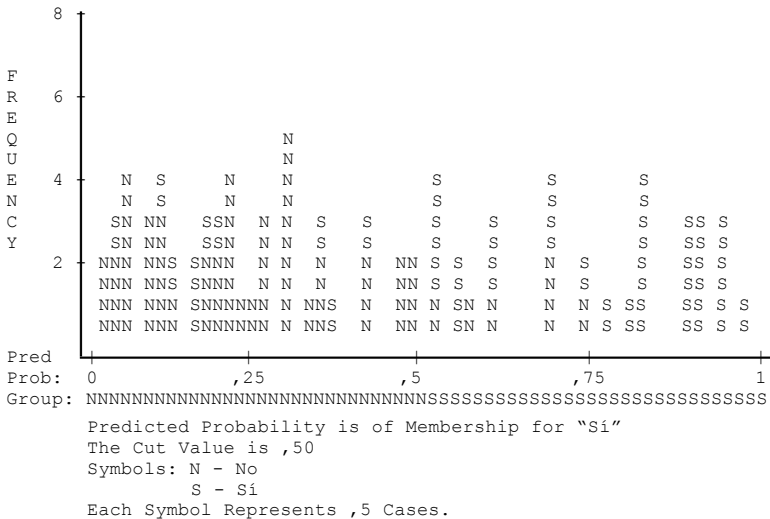
Puesto que las probabilidades pronosticadas toman valores comprendidos entre 0 y 1, para construir la tabla de clasificación a partir de esas probabilidades es imprescindible establecer un **punto de corte**. Los casos con probabilidades pronosticadas mayores que el punto de corte se clasifican en el grupo al que corresponde el código interno 1 (en nuestro ejemplo, el grupo de los pacientes que se recuperan) y los sujetos con probabilidades pronosticadas iguales o menores que el punto de corte se clasifican en el grupo al que corresponde el código interno 0 (en nuestro ejemplo, el grupo de los pacientes que no se recuperan). En el SPSS, este punto de corte es, por defecto, 0,5 (se indica en una nota a pie de tabla), pero este valor puede cambiarse para modificar los porcentajes de clasificación correcta.

Para encontrar el punto de corte óptimo, es decir, la probabilidad pronosticada con la que se consigue la mejor clasificación posible, pueden seguirse diferentes caminos. Uno de ellos consiste en generar múltiples **tablas de clasificación** variando en cada una de ellas el punto de corte hasta maximizar el porcentaje de casos correctamente clasificados (esta tarea puede automatizarse aplicando el procedimiento **Curva COR** —*curva característica de operación del receptor*—; ver Hanley y McNeil, 1982). También puede ayudar bastante a encontrar el punto de corte óptimo un **gráfico de clasificación**. La Figura 5.5 muestra el gráfico de clasificación correspondiente a nuestro ejemplo (este

gráfico puede obtenerse en el SPSS con la opción **Gráfico de clasificación** del subcuadro de diálogo *Regresión logística binaria: Opciones*). Los casos están identificados por una letra; la base del gráfico incluye una leyenda que informa de los símbolos utilizados para diferenciar los casos (N = “no recuperado”, S = “sí recuperado”), del número de casos que representa cada símbolo (*each symbol represents 0,5 cases*) y del punto de corte utilizado (*the cut value is 0,50*). Justo debajo del eje de abscisas está marcado el *territorio* que corresponde a cada pronóstico (la secuencia de símbolos del territorio cambia en el valor del punto de corte). En una situación ideal (clasificación perfecta), todos los símbolos del interior del gráfico estarían ubicados en la vertical de su propio territorio. Los casos que no se encuentran en la vertical de su territorio son casos mal clasificados.

Los casos situados en torno al punto de corte pueden dar pistas acerca de si es posible mejorar la clasificación moviendo el punto de corte. En nuestro ejemplo no parece que esto sea posible. Desplazar el punto de corte hacia la izquierda implicaría incluir rápidamente varios casos N en el grupo de los casos S ; y moverlo hacia la derecha implicaría incluir rápidamente varios casos S en el grupo de los casos N .

Figura 5.5. Gráfico de clasificación basado en las probabilidades pronosticadas



Covariables categóricas

Las variables dicotómicas pueden utilizarse como covariables en un modelo de regresión logística sin ningún tipo de consideración adicional. De hecho, en los ejemplos que hemos utilizado en los apartados anteriores ya hemos trabajado con covariables dicotómicas como *sexo* y *tto*. Con este tipo de variables no existen problemas de estimación ni de interpretación.

Justamente las variables dicotómicas son la solución al problema de cómo incluir variables categóricas politómicas en una ecuación de regresión. Una variable politómica con K categorías puede expresarse, sin pérdida de información, como $K - 1$ variables dicotómicas. A estas variables se les suele llamar variables *dummy* (ficticias). Nosotros seguiremos llamándolas *dicotómicas*. Por ejemplo, la variable *régimen* (régimen hospitalario) de nuestro archivo *Tratamiento adicción alcohol*, que tiene $K = 3$ categorías, puede convertirse en $K - 1 = 2$ variables dicotómicas creando las variables *régimen_1* (con código 1 para el régimen 1 y código 0 para los regímenes 2 y 3) y *régimen_2* (con código 1 para el régimen 2 y código 0 para los regímenes 1 y 3). La Tabla 5.21 recoge este esquema de codificación.

Tabla 5.21. Esquema de codificación *indicador* (para convertir una variable politómica con K categorías en $K - 1$ variables dicotómicas con la misma información)

<i>régimen</i>	<i>régimen_1</i>	<i>régimen_2</i>
1 = interno	1	0
2 = externo	0	1
3 = domiciliario	0	0

Las variables *régimen_1* y *régimen_2*, tomadas juntas, contienen exactamente la misma información que la variable *régimen*. El régimen *interno* queda identificado con el código 1 en *régimen_1* y el código 0 en *régimen_2*; el régimen *externo*, con el código 0 en *régimen_1* y el código 1 en *régimen_2*; y el régimen *domiciliario*, con el código 0 tanto en *régimen_1* como en *régimen_2*. No es necesario crear una tercera variable para identificar el régimen 3 (sería redundante), como tampoco es necesario crear dos variables, sino solo una, para identificar las dos categorías de una variable dicotómica.

Esta forma concreta de convertir una variable politómica en $K - 1$ variables dicotómicas es solamente una entre varias posibles. El SPSS permite elegir entre siete esquemas de codificación distintos, a los que llama *contrastes* (*indicador*, *simple*, *diferencia*, *Helmert*, *repetido*, *polinómico* y *desviación*). Cada uno de estos contrastes responde a una forma concreta de comparación entre las categorías de la variable (la ayuda del programa incluye una aclaración del significado de estos contrastes).

Cualquiera que sea el esquema de codificación aplicado, una vez que una variable politómica ha sido convertida en $K - 1$ variables dicotómicas, ya puede incluirse como covariable en un modelo de regresión.

Veamos como ajustar e interpretar un modelo de regresión logística con *recuperación* como variable dependiente y *régimen* como covariable:

- En el cuadro de diálogo principal, trasladar la variable *recuperación* al cuadro **Dependiente**, la variable *régimen* (régimen hospitalario) a la lista **Covariables** y pulsar el botón **Categórica** para acceder al subcuadro de diálogo *Regresión logística: Definir variables categóricas*.

- Trasladar la variable *régimen* a lista **Covariables categóricas** y, manteniéndola seleccionada, marcar la opción **Primera** como categoría de referencia y pulsar el botón **Cambiar** para hacer efectivo el cambio. Dejar **Indicador**¹⁴ como opción del recuadro **Contraste** y pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestran las Tablas 5.22 y 5.23 (solo explicaremos los resultados relacionados con el hecho de haber incluido una variable categórica en el análisis).

La Tabla 5.22 recoge el esquema de codificación utilizado con la covariable *régimen*. Se han creado dos variables *dicotómicas* (identificadas por las columnas encabezadas 1 y 2). A todas las categorías de la variable *régimen*, excepto a la primera, se les ha asignado el código 1 en la columna correspondiente al parámetro que la va a representar en las estimaciones del modelo. El resto de valores en la misma fila y columna son ceros. Esta información sirve para saber que, más adelante, la categoría *externo* va a estar representada por el parámetro o coeficiente 1 y la categoría *domiciliario* por el parámetro o coeficiente 2. La categoría de referencia, *interno*, tiene ceros en las dos nuevas variables (esta codificación se diferencia de la propuesta en la Tabla 5.21 en que allí se ha tomado, como categoría de referencia, no la primera categoría, sino la última).

Tabla 5.22. Esquema de codificación tipo *indicador*. Variable codificada: *régimen hospitalario*

		Frecuencia	Codificación de parámetros	
			(1)	(2)
Régimen hospitalario	Interno	29	,000	,000
	Externo	30	1,000	,000
	Domiciliario	25	,000	1,000

La Tabla 5.23 ofrece las estimaciones de los coeficientes del modelo y su significación estadística. Estos coeficientes corresponden a la variable *régimen* y a las dos variables dicotómicas creadas en la Tabla 5.22. La tabla también incluye la *constante* del modelo. La ecuación de regresión queda de la siguiente manera:

$$\text{logit}(\text{recuperación} = 1) = -1,34 + 1,75 (\text{régimen}_1) + 1,26 (\text{régimen}_2)$$

La primera fila, encabezada con el nombre de la variable *régimen*, ofrece un contraste del efecto de esa variable. Si este contraste no fuera significativo, carecería de sentido seguir inspeccionando los contrastes (variables dicotómicas) en los que se ha descompuesto su efecto. Puesto que el nivel crítico (*sig.* = 0,011) es menor que 0,05, podemos concluir que la variable *régimen* está relacionada con la *recuperación*.

¹⁴ Para cambiar el tipo de contraste que se desea aplicar a una variable: (1) seleccionar, en la lista **Covariables categóricas**, la covariable categórica cuyo esquema de codificación se desea cambiar (es posible seleccionar un conjunto de covariables para cambiar el tipo de contraste a todas ellas simultáneamente); (2) desplegar el menú **Contraste** para obtener una lista de todos los contrastes disponibles y seleccionar de la lista el contraste deseado; (3) cambiar la categoría de referencia a **Última** o **Primera** según convenga (puede utilizarse la sintaxis para definir una categoría de referencia distinta); (4) pulsar el botón **Cambiar** para actualizar las elecciones hechas.

A continuación aparecen las estimaciones de los coeficientes de regresión y su significación. Un coeficiente significativo ($sig. < 0,05$) indica que la categoría a la que representa difiere significativamente de la categoría de referencia. Las dos categorías representadas, *régimen(1)* y *régimen(2)*, difieren significativamente de la categoría de referencia ($sig. = 0,003$ en el primer caso y $sig. = 0,038$ en el segundo).

Para interpretar estos coeficientes hay que tener en cuenta el esquema de codificación aplicado. En el ejemplo hemos aplicado un esquema de codificación tipo *indicador*. La categoría referencia es *interno*; *régimen(1)* representa a la categoría *externo*; y *régimen(2)* representa a la categoría *domiciliario*. Por tanto, la proporción de recuperaciones entre los sujetos que siguen un régimen hospitalario externo (primera variable dicotómica) difiere de la proporción de recuperaciones entre los sujetos que siguen un régimen hospitalario interno (categoría de referencia). El signo positivo del coeficiente (1,75) indica que la proporción de recuperaciones es mayor en la categoría representada por la primera variable dicotómica (régimen externo) que en la categoría de referencia (régimen interno). Y el valor exponencial del coeficiente, es decir, la *odds ratio* = 5,75, indica que la *odds* de recuperarse con el régimen externo es 5,75 veces la *odds* de recuperarse con el régimen interno.

La proporción de recuperaciones con el régimen domiciliario (categoría representada por la segunda variable dicotómica) difiere de la proporción de recuperaciones con el régimen interno (categoría de referencia). El signo positivo del coeficiente (1,26) indica que la proporción de recuperaciones es mayor en la categoría representada por la segunda variable dicotómica (régimen domiciliario) que en la categoría de referencia (régimen interno). Y el valor exponencial del coeficiente, *odds ratio* = 3,54, indica que la *odds* de recuperarse con el régimen domiciliario es 3,54 veces la *odds* de recuperarse con el régimen interno.

Tabla 5.23. Variables incluidas en la ecuación (estimaciones y significación de los coeficientes)

	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 régimen			8,95	2	,011	
régimen(1)	1,75	,59	8,77	1	,003	5,75
régimen(2)	1,26	,61	4,31	1	,038	3,54
Constante	-1,34	,46	8,59	1	,003	,26

Interacción entre covariables

En los modelos de regresión logística utilizados hasta ahora hemos asumido que las covariables no interactúan; es decir, hemos utilizado modelos que estiman el *logit* de *Y* combinando las covariables aditivamente (sumándolas). Esto implica asumir que, por cada unidad que aumenta una covariable, el modelo de regresión pronostica para el *logit* de *Y* un cambio constante, siempre el mismo, independientemente del valor concreto que tomen el resto de covariables presentes en la ecuación.

Por ejemplo, en el modelo de regresión propuesto en [5.15] se está asumiendo que el *logit* de recuperarse con el tratamiento combinado es 1,84 veces el de recuperarse con el tratamiento estándar *tanto en hombres como en mujeres y cualquiera que sea el número de años de consumo*.

Si la relación entre el *logit* de Y (la variable dependiente) y una determinada covariable (por ejemplo, *tto*) dependiera de los valores de una tercera covariable (por ejemplo, *sexo*), entonces el modelo aditivo no sería un modelo apropiado. Si dos covariables interaccionan, el modelo de regresión debe incluir un término adicional para reflejar esa circunstancia¹⁵.

La forma de incorporar a un modelo de regresión el efecto debido a la interacción entre covariables consiste en incluir el *producto* de las covariables que interaccionan. Un modelo de regresión *no aditivo*, con dos covariables, adopta la siguiente forma:

$$\text{logit}(Y=1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \quad [5.17]$$

Para ajustar con el SPSS un modelo de regresión logística no aditivo con la variable Y como variable dependiente y las variables X_1 y X_2 como covariables:

- En el cuadro de diálogo principal, trasladar la variable Y al cuadro **Dependiente** y las variables X_1 y X_2 a la lista **Covariables**. Seleccionar las variables X_1 y X_2 en la lista de variables y pulsar el botón **>a*b>** para trasladar la interacción entre X_1 y X_2 a la lista de covariables.

Al incluir en la ecuación un término con la interacción $X_1 X_2$ la situación se complica bastante y el significado de los coeficientes cambia. Para facilitar la explicación vamos a considerar tres escenarios: (1) dos covariables dicotómicas, (2) una covariable dicotómica y una cuantitativa y (3) dos covariables cuantitativas.

Dos covariables dicotómicas

En nuestro ejemplo sobre la *recuperación* (Y) de pacientes con problemas de adicción tenemos dos variables dicotómicas: *tratamiento* (X_1) y *sexo* (X_2). Con estas dos variables, el modelo *no aditivo* de regresión logística, es decir, el modelo que incluye, además de los efectos principales *tto* y *sexo*, el efecto de la interacción *tto* $\times$ *sexo*, adopta la siguiente forma:

$$\text{logit}(\text{recuperación}=1) = \beta_0 + \beta_1(\text{tto}) + \beta_2(\text{sexo}) + \beta_3(\text{tto} \times \text{sexo})$$

La Tabla 5.24 muestra los resultados obtenidos al ajustar este modelo de regresión (respecto de un modelo sin interacción, únicamente cambia la tabla de *variables incluidas en la ecuación*).

¹⁵ Para profundizar en todo lo relativo a la interpretación de las interacciones en un modelo de regresión logística puede consultarse Jaccard (2001).

Tabla 5.24. Variables incluidas en la ecuación (con la interacción *tto* x *sexo*)

	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 <i>tto</i>	3,91	1,22	10,20	1	,001	50,00
<i>sexo</i>	-,14	,80	,03	1	,862	,87
<i>tto</i> by <i>sexo</i>	-2,72	1,37	3,97	1	,046	,07
Constante	-1,20	,66	3,35	1	,067	,30

Las estimaciones que ofrece esta tabla permiten formular el siguiente modelo de regresión logística:

$$\text{logit}(\text{recuperación} = 1) = -1,20 + 3,91(\text{tto}) - 0,14(\text{sexo}) - 2,72(\text{tto} \times \text{sexo})$$

Únicamente la variable *tto* y la interacción *tto* × *sexo* tienen asociados coeficientes de regresión significativamente distintos de cero (*sig.* < 0,05). No obstante, interpretaremos todos los coeficientes del modelo para aclarar su significado.

Para ayudar en la interpretación, la Tabla 5.25 contiene las *odds* de recuperarse en cada combinación *tto* × *sexo* (por ejemplo, el valor 0,261 de la primera casilla es la *odds* de recuperarse –cociente entre el número de recuperados y el de no recuperados– entre los hombres que han recibido el tratamiento estándar). Estas *odds* ayudarán a entender el significado de cada coeficiente de regresión.

Tabla 5.25. *Odds* de recuperarse en cada combinación *tto* x *sexo*

<i>Tratamiento</i>	<i>Sexo</i>	
	Hombres	Mujeres
Estándar	0,261	0,3
Combinado	0,857	15

- **Coficiente $\hat{\beta}_0$.** La constante del modelo es, al igual que en cualquier otro modelo de regresión logística, el *logit* estimado para la recuperación cuando todas las covariables incluidas en el modelo valen cero; en nuestro ejemplo, la constante es el *logit* estimado para las mujeres (*sexo* = 0) que han recibido el tratamiento estándar (*tto* = 0). El valor exponencial del coeficiente ($e^{-1,20} = 0,30$) indica que, entre las mujeres que han recibido el tratamiento estándar, el número de recuperaciones es un 30% del número de no recuperaciones.
- **Coficiente $\hat{\beta}_1$ (*tto*).** Para interpretar los coeficientes de regresión asociados a los efectos principales hay que tener en cuenta que estamos ajustando un modelo no aditivo (un modelo con interacción). El coeficiente asociado a la covariable *tto* recoge el efecto de esa covariable sobre la recuperación cuando *sexo* = 0, es decir, cuando los pacientes son mujeres. El valor exponencial de ese coeficiente ($e^{3,91} = 50$) es el resultado de comparar (dividir), en el grupo de mujeres, la *odds* de recu-

perarse con el tratamiento combinado ($tto = 1$) con la *odds* de recuperarse con el tratamiento estándar ($tto = 0$). Ese valor (esa *odds ratio*) está indicando cómo cambia la constante del modelo al pasar del tratamiento estándar al combinado; en concreto, entre las mujeres, la *odds* de recuperarse con el tratamiento combinado (15; ver Tabla 5.25) es 50 veces mayor que la de recuperarse con el tratamiento estándar (0,30). Efectivamente, $15/0,30 = 50$.

- *Coefficiente $\hat{\beta}_2$ (sexo)*. El valor exponencial del coeficiente estimado para la variable *sexo* ($e^{-0,14} = 0,87$) es la *odds ratio* que compara la *odds* de recuperarse entre los hombres (*sexo* = 1) con la *odds* de recuperarse entre las mujeres (*sexo* = 0) con el tratamiento estándar ($tto = 0$). Indica cómo cambia la constante del modelo al pasar del grupo de mujeres al grupo de hombres: entre los pacientes que reciben el tratamiento estándar, la *odds* de recuperarse entre los hombres (0,261; ver Tabla 5.25) es un 87% de la *odds* de recuperarse entre las mujeres (0,30). Efectivamente, $0,261/0,30 = 0,87$. No obstante, puesto que esta diferencia es no significativa (*sig.* = 0,862), no puede afirmarse que la recuperación con el tratamiento estándar sea distinta en los hombres y en las mujeres.
- *Coefficiente $\hat{\beta}_3$ ($tto \times sexo$)*. Por último, el coeficiente de regresión estimado para el efecto de la interacción vale -2,72 y tiene asociado un nivel crítico significativo (*sig.* = 0,046). Para facilitar la interpretación de este coeficiente, comencemos calculando, separadamente para hombres y mujeres, la *odds ratio* que permite comparar el tratamiento combinado con el estándar (las *odds* necesarias para realizar estos cálculos están en la Tabla 5.25):

$$odds\ ratio\ (combinado/est\andar) \mid\ hombres = 0,857/0,261 = 3,284$$

$$odds\ ratio\ (combinado/est\andar) \mid\ mujeres = 15/0,30 = 50$$

Si no existiera efecto de la interacción, estas dos *odds ratios* serían iguales excepto en la parte atribuible a la variabilidad propia del azar muestral. Una diferencia importante entre ambas *odds ratios* estaría indicando que la diferencia entre los dos tratamientos no es la misma entre los hombres y entre las mujeres; o, de forma equivalente, que la diferencia en la recuperación de los hombres y de las mujeres no es la misma con los dos tratamientos.

El cociente entre estas dos *odds ratios* vale $3,284/50 = 0,07$, que es justamente el valor exponencial del coeficiente de regresión correspondiente a la interacción $tto \times sexo$ (ver Tabla 5.24). Este resultado indica que la diferencia entre la *odds* de recuperarse con el tratamiento combinado y la *odds* de recuperarse con el tratamiento estándar no es la misma en los hombres y en las mujeres; en concreto, la *odds ratio* en los hombres (3,284) es únicamente el 7% de esa misma *odds ratio* en las mujeres (50,0).

Por tanto, en los hombres, el tratamiento combinado tiene un beneficio sobre el estándar: con el combinado se recuperan más pacientes. Este beneficio se ha cuantificado con un número (3,284) que indica que, entre los hombres, la *odds* de recuperarse con el tratamiento combinado es aproximadamente el triple de la *odds*

de recuperarse con el tratamiento estándar. Entre las mujeres, el tratamiento combinado también tiene un beneficio sobre el estándar. Ese beneficio se ha cuantificado con un número (50) que indica que, entre las mujeres, la *odds* de recuperarse con el tratamiento combinado es cincuenta veces la *odds* de recuperarse con el tratamiento estándar. El valor exponencial del coeficiente ($e^{-2,72} = 0,07$) está indicando esta diferencia entre hombres y mujeres: 3,284 es un 7% de 50.

Si la variable *sexo* se hubiera codificado al revés (es decir, asignando el código 1 a las mujeres y el código 0 a los hombres), el cociente entre las correspondientes *odds ratios* también se habría calculado la revés: $50/2,284 = 15,23$. Este resultado sigue indicando que la diferencia entre las *odds* de recuperarse con el tratamiento combinado y con el estándar no es la misma en hombres y en mujeres; en concreto, el cociente entre esas *odds* (la *odds ratio*) en el grupo de mujeres (50) es 15,23 veces mayor que ese mismo cociente en el grupo de hombres (3,284).

Una covariable dicotómica y una cuantitativa

Consideremos ahora un modelo de regresión logística *no aditivo* con la variable *recuperación* (Y) como variable dependiente y las variables *tratamiento* (X_1) y *años consumiendo* (X_2) como covariables:

$$\text{logit}(\text{recuperación} = 1) = \beta_0 + \beta_1(\text{tto}) + \beta_2(\text{años}_c) + \beta_3(\text{tto} \times \text{años}_c)$$

Para facilitar la interpretación de los coeficientes de regresión, en lugar de la variable original *años* (años consumiendo) estamos utilizando la variable *años_c* (años consumiendo *centrada*), la cual se ha centrado restando 14 puntos (que es el valor de la mediana) a todas las puntuaciones de la variable *años*; al aplicar esta transformación, el valor *años_c* = 0 se refiere a los pacientes con 14 años de consumo.

La Tabla 5.26 muestra los resultados del análisis. Las estimaciones que ofrece la tabla permiten construir el siguiente modelo de regresión:

$$\text{logit}(\text{recuperación} = 1) = -1,31 + 1,86(\text{tto}) + 0,01(\text{años}_c) - 0,54(\text{tto} \times \text{años}_c)$$

Tabla 5.26. Variables incluidas en la ecuación (con la interacción *tto* x *años*)

	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 tto	1,86	,58	10,19	1	,001	6,43
años_c	,01	,09	,02	1	,886	1,01
tto by años_c	-,54	,19	7,92	1	,005	,58
Constante	-1,31	,39	11,46	1	,001	,27

- *Coficiente $\hat{\beta}_0$* . La constante del modelo es el *logit* estimado para la recuperación cuando ambas covariables, *tto* y *años*, valen cero; en nuestro ejemplo, la constante es el *logit* estimado para los pacientes que reciben el tratamiento estándar (*tto* = 0) y que llevan 14 años consumiendo (*años_c* = 0). El valor exponencial del coeficien-

te ($e^{-1,31} = 0,27$) indica que, entre los pacientes que han recibido el tratamiento estándar tras llevar 14 años consumiendo, el número de recuperaciones es un 27% del número de no recuperaciones.

- **Coefficiente $\hat{\beta}_1$ (*tto*).** Para interpretar los coeficientes de regresión asociados a los efectos principales hay que tener en cuenta que se trata de un modelo que incluye una interacción significativa. El coeficiente estimado para la covariable *tto* (1,86) recoge el efecto de esa covariable cuando *años_c* = 0, es decir, cuando los pacientes llevan consumiendo 14 años. Por tanto, su valor exponencial ($e^{1,86} = 6,43$) es la *odds ratio* que compara, en los pacientes con 14 años de consumo, la *odds* de recuperarse con el tratamiento combinado (*tto* = 1) con la *odds* de recuperarse con el tratamiento estándar (*tto* = 0). El valor de esta *odds ratio* indica que, entre los pacientes con 14 años de consumo, la *odds* de recuperarse con el tratamiento combinado es 6,43 veces la de recuperarse con el tratamiento estándar.
- **Coefficiente $\hat{\beta}_2$ (*años_c*).** El valor exponencial del coeficiente de regresión asociado a la covariable *años_c* ($e^{0,01} = 1,01$) es el valor por el que queda multiplicada la *odds* de recuperarse con el tratamiento estándar (*tto* = 0) con cada año más de consumo. Por tanto, 1,01 indica que, entre los pacientes que reciben el tratamiento estándar, la *odds* de recuperarse va aumentando un 1% con cada año más de consumo. No obstante, como este incremento del 1% es no significativo (*sig.* = 0,886), no puede concluirse que la recuperación con el tratamiento estándar cambie con los años de consumo.
- **Coefficiente $\hat{\beta}_3$ (*tto* × *años_c*).** Por último, el coeficiente de regresión estimado para el efecto de la interacción vale -0,54 y tiene asociado un nivel crítico significativo (*sig.* = 0,005). El valor exponencial de este coeficiente ($e^{-0,54} = 0,58$) indica cómo cambia la diferencia en la recuperación con ambos tratamientos al aumentar los años de consumo. En concreto, 0,58 indica que el cociente entre la *odds* de recuperarse con el tratamiento combinado y la *odds* de recuperarse con el tratamiento estándar disminuye un 42% con cada año más de consumo.

En esta interpretación del efecto de la interacción hemos puesto el énfasis en la relación entre la variable *tto* y la variable dependiente *recuperación*; es decir, hemos considerado que la variable *años_c* (la variable cuantitativa) desempeña el rol de variable *moderadora* de la relación entre los *tratamientos* y la *recuperación*. Los mismos datos pueden interpretarse tomando la variable *tto* (la variable categórica) como moderadora y, por tanto, poniendo el énfasis de la interpretación en la relación entre los *años de consumo* y la *recuperación*.

El valor exponencial del coeficiente estimado para la variable *años_c* vale 1,01 cuando se aplica el tratamiento estándar y 0,59 cuando se aplica el combinado (este valor puede obtenerse intercambiando los códigos 1 y 0 asignados a los tratamientos). Por tanto, la relación entre la recuperación y los años de consumo no parece ser la misma con ambos tratamientos: con el estándar se estima que la relación aumenta un 1% con cada año de consumo; con el combinado se estima que la relación disminuye un 41% con cada año de consumo (la relación entre la *recuperación* y

los *años de consumo* es más intensa cuando se aplica el tratamiento combinado). El cociente entre ambos exponentes es justamente el coeficiente estimado para el efecto de la interacción: $0,59/1,01 = 0,58$. Por tanto, 0,58 es el factor por el que queda multiplicada la *odds ratio* que relaciona la *recuperación* con los *años de consumo* cuando se cambia del tratamiento estándar al combinado.

Dos covariables cuantitativas

Para terminar con la explicación del efecto de la interacción, consideremos un modelo *no aditivo* con la *recuperación* (Y) como variable dependiente y las variables *edad* (X_1) y *años consumiendo* (X_2) como covariables:

$$\text{logit}(\text{recuperación} = 1) = \beta_0 + \beta_1(\text{edad}_c) + \beta_2(\text{años}_c) + \beta_3(\text{edad}_c \times \text{años}_c)$$

Puesto que ambas covariables son cuantitativas, hemos decidido centrarlas en la mediana (34 para la edad y 14 para los años de consumo). Por tanto, el valor $\text{edad}_c = 0$ se refiere a 34 años de edad y el valor $\text{años}_c = 0$ se refiere a 14 años de consumo (recordemos que las covariables cuantitativas se centran únicamente para facilitar la interpretación de los coeficientes de regresión).

La Tabla 5.27 muestra los resultados del análisis. Con las estimaciones que ofrece la tabla se obtiene el siguiente modelo de regresión:

$$\text{logit}(\text{rec.} = 1) = -0,30 + 0,09(\text{edad}_c) - 0,31(\text{años}_c) - 0,01(\text{edad}_c \times \text{años}_c)$$

Tabla 5.27. Variables incluidas en la ecuación (con la interacción *edad* x *años*)

	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1						
edad_c	,09	,05	3,57	1	,059	1,10
años_c	-,31	,09	13,10	1	,000	,73
edad_c by años_c	-,01	,01	1,11	1	,293	,99
Constante	-,30	,28	1,15	1	,284	,74

- **Coeficiente $\hat{\beta}_0$.** La constante del modelo es el *logit* estimado para la recuperación cuando todas las covariables valen cero. En nuestro ejemplo, el *logit* estimado para los pacientes con 34 años de edad y 14 de consumo. El valor exponencial de la constante ($e^{-0,30} = 0,74$) indica que, en esos pacientes, el número de recuperaciones es un 74% del de no recuperaciones. No obstante, como el valor estimado para la constante no alcanza la significación estadística (*sig.* = 0,284), no puede afirmarse que, en esos pacientes, la recuperación se dé menos que la no recuperación.
- **Coeficiente $\hat{\beta}_1$ (*edad_c*).** Si el efecto de la interacción fuera significativo, el coeficiente asociado a la covariable *edad_c* estaría reflejando el efecto de esa covariable cuando *años_c* = 0 (14 años de consumo), pero como el efecto de la interacción es no significativo, el efecto de *edad_c* debe referirse a todos los pacientes. Por tanto, el valor exponencial del coeficiente ($e^{0,09} = 1,10$) indica que, cualesquiera que

sean los años de consumo, la *odds* de recuperarse aumenta un 10% cuando la edad aumenta un año. Pero, dado que el valor del coeficiente no alcanza la significación estadística (*sig.* = 0,059), lo razonable es concluir que no existe evidencia de que la edad esté relacionada con la recuperación.

- *Coefficiente $\hat{\beta}_2$ (*años_c*)*. El valor exponencial del coeficiente de regresión asociado a la covariable *años_c* ($e^{-0,31} = 0,73$) es el valor por el que queda multiplicada la *odds* de recuperarse con cada año más de consumo cuando *edad_c* = 0 (34 años). Ahora bien, como el efecto de la interacción es no significativo, el valor 0,73 se refiere a todos los pacientes, no solo a los que tienen 34 años. Por tanto, la *odds* de recuperarse va disminuyendo un 27% con cada año más de consumo.
- *Coefficiente $\hat{\beta}_3$ (*edad_c* × *años_c*)*. Por último, el coeficiente de regresión estimado para el efecto de la interacción vale -0,01. Su valor exponencial ($e^{-0,01} = 0,99$) indica cómo cambia la relación entre la recuperación y los años de consumo al ir aumentando la edad. En concreto, 0,99 indica que la *odds ratio* que relaciona la recuperación con los años de consumo disminuye un 1% con cada año más de edad. Pero este cambio no alcanza la significación estadística (*sig.* = 0,293).

Regresión logística jerárquica o por pasos

Hasta ahora hemos asumido en todo momento que la decisión sobre qué variables debe incluir una ecuación de regresión es responsabilidad del investigador. Es decir, hemos asumido que es el propio investigador quien, generalmente guiado por una hipótesis de trabajo, decide con qué variables va a construir su ecuación.

Sin embargo, no es infrecuente encontrar situaciones en las que no existe una hipótesis de trabajo que oriente al investigador en la elección de las variables realmente relevantes. En estos casos se podría comenzar incluyendo en la ecuación todas las variables que se intuye que pueden contribuir a entender o explicar el fenómeno estudiado para, a continuación, eliminar las variables con coeficientes de regresión no significativos. Pero esta estrategia es bastante problemática: dado que los coeficientes de regresión son coeficientes *parciales* (pues el valor de un coeficiente de regresión depende del resto de coeficientes presentes en la ecuación), eliminar más de una variable al mismo tiempo impide valorar el comportamiento individual de las variables eliminadas.

Es preferible proceder *jerárquicamente* o *por pasos*, tal como hemos hecho ya al estudiar la regresión lineal (ver el Capítulo 10 del segundo volumen). Con la regresión por pasos se pretende encontrar el modelo capaz de ofrecer el mejor ajuste posible con el menor número de variables. Con esta forma de proceder se intenta hacer compatibles los dos principios que deben guiar la elección de un modelo estadístico: (1) el principio de **parsimonia**, según el cual un modelo estadístico debe incluir el menor número posible de variables para facilitar la interpretación de los resultados y hacer el modelo más generalizable y (2) el principio de **máximo ajuste**, según el cual un modelo estadístico debe conseguir explicar lo mejor posible el comportamiento de la variable dependiente.

Existen varias estrategias para seleccionar las covariables que deben formar parte del modelo final: (1) la inclusión forzosa, (2) la selección por pasos y (3) la selección por bloques:

1. La estrategia de **inclusión forzosa** construye el modelo de regresión con todas las covariables seleccionadas. Esta estrategia tiene la doble ventaja de que permite valorar el efecto conjunto de todas las covariables elegidas y de que el modelo que se construye contiene las covariables que se consideran relevantes desde el punto de vista teórico. Como contrapartida, suele darse el caso de que el modelo final incluye covariables irrelevantes que no contribuyen al ajuste.
2. La **selección por pasos** utiliza criterios estadísticos para incluir en el modelo final únicamente las covariables que contribuyen al ajuste. La ventaja de esta estrategia es que permite construir modelos que no incluyen variables irrelevantes desde el punto de vista estadístico. El inconveniente es que puede dejar fuera de la ecuación variables teórica o conceptualmente relevantes¹⁶.
3. La **selección por bloques** permite controlar la inclusión/exclusión de bloques de variables. Se puede controlar qué variables se incluyen/excluyen en cada paso (en cada bloque) y el orden en que se debe incluir/excluir cada bloque. La principal ventaja de esta estrategia radica en la posibilidad de comparar modelos jerárquicos o anidados valorando simultáneamente la significación de más de una covariable. A esta estrategia se le suele llamar **regresión jerárquica**¹⁷.

En la selección *por pasos* y en la selección *por bloques* se puede proceder hacia delante o hacia atrás. Los métodos **hacia delante** parten del modelo nulo y van incorporando variables paso a paso hasta que no quedan variables que contribuyan a mejorar su ajuste. Los métodos **hacia atrás** parten del modelo que incluye todas las variables elegidas como posibles covariables y van excluyendo variables paso a paso hasta que solo quedan las que contribuyen significativamente al ajuste.

Al elegir una estrategia de selección *por pasos* o *por bloques*, el SPSS permite construir el modelo de regresión aplicando diferentes métodos de selección de variables. Todos ellos se basan en criterios *estadísticos*: incluyen en el modelo las covariables que contribuyen al ajuste; excluyen las que no contribuyen al ajuste. Para incluir covariables todos los métodos utilizan el *estadístico de puntuación* de Rao. Para excluir covariables se puede elegir entre tres estadísticos: la *razón de verosimilitudes*, el *estadístico de Wald* y el *estadístico condicional* (ver Lawless y Singhal, 1978). La *razón de verosimilitudes*

¹⁶ Construir una ecuación de regresión por pasos no siempre resulta ser una idea tan buena como en principio podría parecer. Si el objetivo del análisis es efectuar pronósticos y no existe una hipótesis de trabajo que justifique la elección de unas covariables u otras, proceder jerárquicamente o por pasos puede resultar una estrategia válida porque se consigue el máximo ajuste con el menor número de covariables. Si el objetivo del análisis es obtener evidencia empírica sobre alguna hipótesis de trabajo, entonces proceder por pasos podría resultar más perjudicial que beneficioso, pues podría ocurrir que el modelo con el mejor ajuste incluyera variables teóricamente irrelevantes y que el ajuste de ese modelo fuera solo ligeramente mejor que el de un modelo con variables teóricamente relevantes (ver Henderson y Denison, 1989, o Huberty, 1989).

¹⁷ Esta estrategia de construcción de un modelo de regresión por bloques de variables no debe confundirse con la regresión *multinivel* (ver capítulo anterior), la cual, a veces, también recibe el nombre de regresión *jerárquica*.

suele ser el estadístico más recomendado, pero el estadístico *condicional* es computacionalmente más eficiente.

Veamos cómo realizar un análisis de regresión logística utilizando un método de selección por pasos. Seguimos utilizando el archivo *Tratamiento adicción alcohol*:

- En el cuadro de diálogo principal, trasladar la variable *recuperación* al cuadro **Dependiente** y las variables *sexo*, *edad*, *años* (años de consumo) y *tto* (tratamiento) a la lista de **Covariables**.
- En el menú desplegable del recuadro **Método** seleccionar el método **Adelante: RV**.

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestran las Tablas 5.28 a 5.33. Los estadísticos de *puntuación* de la Tabla 5.28 indican lo que ocurriría con cada covariable de ser ella la elegida en el primer paso. La covariable elegida en este paso es la que tiene asociado el estadístico de puntuación más alto al tiempo que un nivel crítico menor que 0,05. En nuestro ejemplo, *tto*.

El estadístico de puntuación de la última fila (*estadísticos globales*) permite contrastar la hipótesis de que todos los coeficientes de regresión, excluida la constante, valen cero en la población. Si no puede rechazarse esta hipótesis, no podrá construirse un modelo que mejore el ajuste del modelo nulo.

Tabla 5.28. Variables no incluidas en la ecuación en el paso 0

			Puntuación	gl	Sig.
Paso 0	Variables	sexo	6,68	1	,005
		edad	,37	1	,545
		años	12,69	1	,000
		tto	15,75	1	,000
	Estadísticos globales		29,42	4	,000

La Tabla 5.29 ofrece una prueba de ajuste global. El estadístico que aparece con el nombre *chi-cuadrado* es la razón de verosimilitudes G^2 (ver ecuación [5.9]). Este estadístico permite contrastar, en cada paso, la hipótesis nula de que el modelo propuesto en ese paso no mejora el ajuste (o no reduce el desajuste) del modelo nulo. La tabla informa de las variaciones producidas en el desajuste como consecuencia de la incorporación (o eliminación) de cada nueva variable.

En cada paso se muestran tres valores: (1) *paso* muestra el cambio que se produce en la desviación entre un paso y el siguiente; permite contrastar la hipótesis de que el efecto de la variable incluida en un determinado paso es nulo; (2) *bloque* recoge el cambio que se produce en la desviación entre un bloque y el siguiente cuando se solicita el ajuste de varios modelos formados por distintos bloques de variables; permite contrastar la hipótesis de que el efecto asociado a cada bloque de variables es nulo (esta información únicamente es útil si se utiliza un método de selección de variables por bloques); y (3) *modelo* informa del cambio que se produce en la desviación entre el modelo nulo (paso 0) y el modelo construido en cada paso.

Tabla 5.29. Pruebas *omnibus* sobre los coeficientes del modelo (contrastes de ajuste global)

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	16,34	1	,000
	Bloque	16,34	1	,000
	Modelo	16,34	1	,000
Paso 2	Paso	9,57	1	,002
	Bloque	25,90	2	,000
	Modelo	25,90	2	,000
Paso 3	Paso	5,40	1	,020
	Bloque	31,30	3	,000
	Modelo	31,30	3	,000

En el primer paso se ha elegido la variable *tto* (ver Tabla 5.31); su incorporación representa una reducción significativa del desajuste del modelo nulo (*chi-cuadrado* = 16,34; *sig.* < 0,0005). En el segundo paso se ha elegido la variable *años* (ver Tabla 5.31); su incorporación (*paso*) supone una reducción significativa del desajuste del modelo del paso anterior (*chi-cuadrado* = 9,57; *sig.* = 0,002), y el modelo resultante (*modelo*), que en este segundo paso incluye las covariables *tto* y *años*, permite reducir significativamente el desajuste del modelo nulo (*chi-cuadrado* = 25,90; *sig.* < 0,0005). En el tercer paso se ha elegido la variable *sexo* (ver Tabla 5.31); su incorporación (*paso*) supone una reducción significativa del desajuste del modelo del paso anterior (*chi-cuadrado* = 5,40; *sig.* = 0,020), y el modelo resultante (*modelo*), que en este tercer paso incluye las covariables *tto*, *años* y *sexo*, permite reducir significativamente el desajuste del modelo nulo (*chi-cuadrado* = 31,30; *sig.* < 0,0005).

El ajuste por pasos se detiene en el tercer paso. La covariable *edad* queda fuera del modelo porque incluirla no contribuye a reducir el desajuste del modelo del tercer paso. Tal como cabía esperar, en el último paso es donde el estadístico G^2 toma su valor más alto, decir, donde se consigue la mayor reducción del desajuste del modelo nulo.

En los estadísticos de ajuste global de la Tabla 5.30 también se puede apreciar que el desajuste se va reduciendo en cada paso: el valor de la desviación $-2LL$ (en la tabla, $-2 \log de la verosimilitud$) va disminuyendo y los estadísticos tipo R^2 van aumentando. El estadístico de Nagelkerke indica que el modelo final, es decir, el modelo que incluye las covariables *tto*, *años* y *sexo*, consigue reducir en un 42 % el desajuste del modelo nulo.

La Tabla 5.31 contiene los modelos de regresión que se han ido construyendo en cada paso. El último paso es, por lo general, el paso en el que conviene centrarse, pues es el que contiene el modelo final. De las cuatro covariables elegidas para el análisis, el método de selección por pasos se ha quedado con tres: *tto*, *años* y *sexo*. La variable *edad* ha quedado fuera porque no contribuye a reducir el desajuste. El modelo final solo incluye variables con coeficientes de regresión significativamente distintos de cero. Los coeficientes de regresión y los contrastes que contiene esta tabla se interpretan tal como ya hemos hecho a propósito de la Tabla 5.20 (el modelo final es idéntico al obtenido allí).

Tabla 5.30. Resumen de los modelos (estadísticos de ajuste global)

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	98,39	,18	,24
2	88,82	,27	,36
3	83,43	,31	,42

Tabla 5.31. Variables incluidas en el modelo (estimaciones y significación de los coeficientes)

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 ^a	tto	1,89	,50	14,53	1	,000	6,60
	Constante	-1,30	,38	11,94	1	,001	,27
Paso 2 ^b	años	-,18	,06	8,09	1	,004	,83
	tto	1,77	,53	11,26	1	,001	5,86
Paso 3 ^c	Constante	1,29	,95	1,83	1	,176	3,63
	sexo	-1,33	,59	5,06	1	,024	,26
	años	-,18	,07	7,34	1	,007	,84
	tto	1,84	,55	11,01	1	,001	6,27
	Constante	2,11	1,07	3,86	1	,049	8,23

a. Variable(s) introducida(s) en el paso 1: tto.

b. Variable(s) introducida(s) en el paso 2: años.

c. Variable(s) introducida(s) en el paso 3: sexo.

La Tabla 5.32 informa de lo que ocurriría en cada paso con cada una de las covariables ya incluidas en el modelo si se decidiera expulsarlas del mismo. Aunque los métodos de selección de variables por pasos *hacia delante* funcionan incluyendo una covariable en cada paso, también permiten excluir una variable previamente incluida si el correspondiente coeficiente de regresión deja de ser significativo como consecuencia de la incorporación de nuevas variables.

La columna encabezada *cambio en -2 log de la verosimilitud* contiene la razón de verosimilitudes G^2 . Recordemos que este estadístico sirve para comparar las desviaciones de dos modelos jerárquicos. Aquí sirve para valorar, en cada paso, el cambio que se produce en la desviación del modelo al eliminar cada una de las variables que incluye. Por ejemplo, 16,34 es el cambio (aumento) que experimentaría la desviación del modelo del paso 1 (el modelo que incluye la covariable *tto*) si se eliminara la covariable *tto*; 9,57 es el cambio (aumento) que experimentaría la desviación del modelo del paso 2 (el modelo que incluye las covariables *tto* y *años*) si se eliminara la covariable *años*; etc. Si el cambio en la desviación tiene asociado un nivel crítico (*sig. del cambio*) menor que 0,05, eliminar la correspondiente covariable supondría un aumento significativo del desajuste. En nuestro ejemplo, en ningún momento se excluye ninguna de las covariables previamente incluidas: cualquier exclusión supondría aumentar el desajuste.

La columna encabezada *log verosimilitud del modelo* ofrece los valores a partir de los cuales se calcula tanto la desviación de cada modelo como el cambio que se va produciendo en la desviación. Por ejemplo, -57,36 multiplicado por -2 (o sea, 114,72) es la desviación del modelo nulo, es decir, la desviación del modelo que se está ajustando

en el paso 1 cuando se elimina del mismo la única covariable que incluye (*tto*). Y el valor $-49,20$ multiplicado por -2 (o sea, $98,40$) es la desviación del modelo que se está ajustando en el paso 2 cuando se elimina del mismo la covariable *años*, es decir, la desviación del modelo que únicamente incluye la covariable *tto* (ver Tabla 5.10). Etc.

Tabla 5.32. Pérdida de ajuste del modelo al excluir variables

Variable	Log verosimilitud del modelo	Cambio en $-2 \log$ de la verosimilitud	gl	Sig. del cambio
Paso 1 <i>tto</i>	-57,36	16,34	1	,000
Paso 2 <i>años</i>	-49,20	9,57	1	,002
<i>tto</i>	-50,59	12,35	1	,000
Paso 3 <i>sexo</i>	-44,41	5,40	1	,020
<i>años</i>	-45,96	8,50	1	,004
<i>tto</i>	-47,86	12,28	1	,000

Finalmente, la Tabla 5.33 muestra información sobre lo que ocurre en cada paso con las variables todavía no incluidas en el modelo. La variable que será incorporada al modelo en el siguiente paso es aquella a la que le corresponde, en el paso previo, el *estadístico de puntuación* más alto (siempre que éste sea significativo). La tabla muestra que, de las variables no incluidas en el primer paso, *años* es la que tiene un *estadístico de puntuación* más alto (9,21); como, además, el correspondiente nivel crítico es significativo (*sig.* = 0,002), *años* es la variable incorporada al modelo en el segundo paso.

En el resto de los pasos se aplica el mismo criterio. En el segundo paso quedan fuera del modelo las variables *sexo* y *edad*. De las dos, *sexo* es a la que le corresponde el *estadístico de puntuación* más alto (5,41) y, además, es la única que tiene asociado un nivel crítico significativo (*sig.* 0,020); por tanto, la variable *sexo* es la elegida en el tercer paso.

En el tercer paso solamente queda fuera del modelo la variable *edad*. Y queda definitivamente fuera porque no contribuye a reducir el desajuste del modelo que incluye las otras tres covariables (*sig.* = 0,074 > 0,05)

Tabla 5.33. Variables no incluidas en el modelo

			Puntuación	gl	Sig.
Paso 1	Variables	<i>sexo</i>	6,42	1	,011
		<i>edad</i>	1,21	1	,270
		<i>años</i>	9,21	1	,002
	Estadísticos globales		16,18	3	,001
Paso 2	Variables	<i>sexo</i>	5,41	1	,020
		<i>edad</i>	1,42	1	,233
	Estadísticos globales		8,36	2	,015
Paso 3	Variables	<i>edad</i>	3,19	1	,074
	Estadísticos globales		3,19	1	,074

Supuestos del modelo de regresión logística

Ya sabemos que, para que un modelo lineal funcione correctamente, es necesario que se den una serie de condiciones (ver, en el Capítulo 1, el apartado *Chequear los supuestos del modelo*). En un modelo de regresión logística estas condiciones son, básicamente, cuatro. Nos referiremos a ellas, abreviadamente, como: (1) linealidad, (2) no-colinealidad, (3) independencia y (4) dispersión proporcional a la media.

Linealidad

El primero y más importante supuesto de un análisis de regresión logística es que el modelo está correctamente especificado. Se comete un **error de especificación** cuando no se eligen bien las variables independientes (bien porque hay otra u otras variables que podrían explicar mejor el comportamiento de la variable dependiente, bien porque se han incluido en el modelo variables irrelevantes) o cuando, habiendo elegido bien las variables independientes, su relación con el *logit* de Y no es de tipo lineal.

En primer lugar, si faltan en el modelo *variables importantes*, no solo el ajuste no será del todo bueno, sino que las estimaciones de los coeficientes estarán sesgadas; y sin una teoría que dirija la búsqueda de nuevas variables, este problema no tiene fácil solución. Si el modelo incluye *variables irrelevantes*, las estimaciones de los coeficientes serán poco eficientes (los errores típicos estarán inflados); pero este problema tiene fácil solución porque las variables irrelevantes suelen detectarse fácilmente a partir de la significación de sus coeficientes.

En segundo lugar, un modelo de regresión logística estima, para el *logit* de Y , un cambio *constante* de tamaño β_j por cada unidad que aumenta X_j (para cualquier combinación entre los valores del resto de covariables). Este cambio *constante* es el que le confiere al modelo su carácter de *lineal*. El supuesto de linealidad es crucial: no tiene sentido utilizar una ecuación lineal si la relación subyacente no es lineal.

El supuesto de linealidad puede contrastarse aplicando diferentes estrategias (ver Harrell, 2001). Una sencilla consiste en dividir la covariable en categorías igualmente espaciadas y estimar los coeficientes de regresión asociados a cada categoría. Si la relación entre el *logit* de Y y la covariable categorizada es lineal, los coeficientes estimados para las categorías deberán aumentar o disminuir de forma aproximadamente lineal.

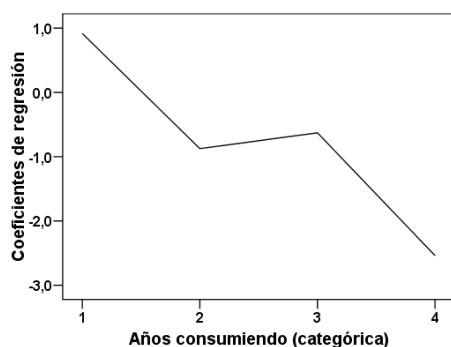
Para aplicar esta estrategia, hemos transformado la variable *años* en una variable categórica, *años_cat*, con puntos de corte en 4, 8, 12 y 16 años, y la hemos incluido en el análisis aplicándole una codificación de tipo *indicador* y fijando la primera categoría como categoría de referencia. Los coeficientes de regresión obtenidos están representados en la Figura 5.6. El gráfico muestra una tendencia básicamente lineal, con un leve escalón que no parece que sea suficiente para alterar la tendencia general.

Esta estrategia tiene su utilidad, pero la valoración que se hace del tamaño de los coeficientes es solo aproximada. Se consigue mayor precisión aplicando contrastes de tipo polinómico. Estos contrastes sirven para estudiar si la relación entre la variable

dependiente y las covariables es lineal, cuadrática, cúbica, etc. (el lector no familiarizado con este tipo de contrastes puede consultar, en el Capítulo 6 del segundo volumen, el apartado *Comparaciones de tendencia*).

Al aplicar estos contrastes con la *recuperación* como variable dependiente y los *años de consumo* (*años_cat*) como covariable, únicamente la tendencia lineal ha resultado significativa (con $p = 0,026$); ninguna de las restantes tendencias ha alcanzado la significación estadística (debe tenerse en cuenta que una categorización distinta de la variable *años* podría arrojar resultados ligeramente diferentes).

Figura 5.6. Coeficientes asociados a las categorías de la variable *años_cat*



Además de asumir que la relación entre el *logit* de Y y el conjunto de covariables es lineal, si la ecuación de regresión no incluye términos referidos a las posibles interacciones entre covariables, también se está asumiendo que el cambio estimado para Y por cada unidad que aumenta X_j es siempre el mismo independientemente del valor concreto que tomen el resto de las covariables incluidas en la ecuación, es decir, independientemente del valor concreto en el que permanezcan constantes el resto de las covariables.

Si la relación entre Y y una determinada X_j depende de los valores que tome alguna otra X_j , entonces el modelo aditivo (el modelo que no incluye interacciones entre covariables) no es un modelo apropiado. En presencia de interacción entre covariables es importante no dejar fuera de la ecuación el producto de las variables que interactúan (ver, más atrás, el apartado *Interacción entre variables independientes*).

No colinealidad

El concepto de *colinealidad* (o *multicolinealidad*) se refiere a la relación entre variables independientes. Existe colinealidad perfecta cuando una variable independiente es función lineal perfecta de otra u otras variables independientes.

Para poder estimar los coeficientes de regresión es imprescindible que no exista colinealidad perfecta pues, si existe, no hay solución única para las estimaciones. La

colinealidad perfecta es infrecuente¹⁸, sin embargo, no es infrecuente que exista cierto grado de colinealidad (es improbable que un conjunto de variables sean completamente independientes). La cuestión, por tanto, no es si existe o no colinealidad, sino si el grado de colinealidad existente representa un problema. El problema de una colinealidad elevada es que infla el tamaño de los errores típicos de los coeficientes. Y esto tiene una doble consecuencia; por un lado, es más difícil rechazar las hipótesis nulas de que los coeficientes de regresión valen cero en la población; por otro, las estimaciones de los coeficientes se vuelven inestables (pequeños cambios en los datos pueden llevar a cambios importantes en las estimaciones).

Existen algunos estadísticos que pueden ayudar a detectar si el grado de colinealidad está causando problemas. El nivel de **tolerancia** de una variable independiente X_j se obtiene restando a 1 el coeficiente de determinación correspondiente a la ecuación de regresión de X_j sobre el resto de variables independientes ($1 - R_j^2$). Un nivel de tolerancia próximo a 1 indica que la variable X_j no está relacionada con el resto de variables independientes; un nivel de tolerancia próximo a 0 indica que la variable X_j está muy relacionada con el resto de variables independientes. Suele considerarse que los problemas asociados a la presencia de elevada colinealidad empiezan con tolerancias menores que 0,10.

A los valores inversos de los niveles de tolerancia, $1/(1 - R_j^2)$, se les llama **factores de inflación de la varianza (FIV_j)**. Reciben este nombre porque reflejan el aumento que experimenta la varianza de cada coeficiente de regresión como consecuencia de la relación existente entre las variables independientes. Los **FIV_j** informan exactamente de lo mismo que los niveles de tolerancia. Valores mayores que 10 suelen ir acompañados de los problemas de estimación asociados a un exceso de colinealidad.

El procedimiento **Regresión logística binaria** no ofrece ni los niveles de tolerancia ni los factores de inflación de la varianza. Pero pueden obtenerse con el procedimiento **Regresión lineal** tal como se ha explicado en el Capítulo 10 del segundo volumen (puesto que al valorar el grado de colinealidad únicamente intervienen las variables independientes, el hecho de que se esté trabajando con una respuesta dicotómica es irrelevante a la hora de diagnosticar la colinealidad). Por lo general, el exceso de colinealidad es un problema más fácil de detectar que de resolver. No obstante, en el Capítulo 10 del segundo volumen se ofrecen algunas soluciones que pueden aplicarse cuando el exceso de colinealidad es un problema.

Independencia

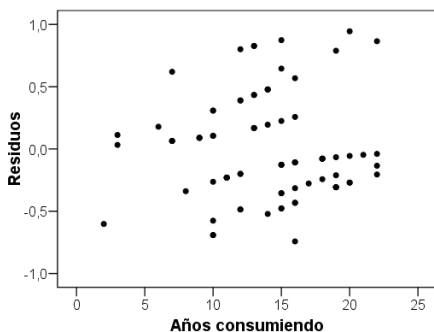
La mayor parte de los procedimientos estadísticos asumen que se está trabajando con observaciones (por tanto, con errores) independientes entre sí. El análisis de regresión logística no es una excepción.

¹⁸ Se da colinealidad perfecta, por ejemplo, cuando se incluye en el análisis una variable que es suma de otras que también se incluyen (como los ítems de una escala y la puntuación total en la escala obtenida como la suma de los ítems), o cuando se incluyen variables cuyos valores suman una constante (como el porcentaje de tiempo libre dedicado a cada una de un conjunto de actividades).

En general, la independencia viene garantizada por el muestreo: si el muestreo es aleatorio, los errores tenderán a mostrar una pauta aleatoria. La independencia entre errores significa que no están autocorrelacionados, es decir, que no aumentan o disminuyen siguiendo una pauta discernible. Este supuesto suele incumplirse en los datos procedentes de estudios longitudinales (como en el caso de las series temporales), en los datos recogidos secuencialmente (donde los terapeutas pueden mejorar su forma de administrar un tratamiento, los sujetos pueden mostrar fatiga, los aparatos pueden sufrir algún tipo de desgaste), en los datos recogidos en grupos homogéneos de sujetos pero diferentes entre sí (grupos de diferente ideología política o religiosa, grupos de diferente clase social), etc. En este tipo de estudios, el error asociado a un caso tiende a parecerse a los errores de los casos adyacentes; y esta circunstancia suele producir *sobredispersión* (ver siguiente apartado).

El supuesto de independencia también se refiere a las covariables. Puesto que los errores representan la parte de Y que el modelo de regresión no consigue explicar, es razonable esperar que no estén relacionados con las covariables incluidas en la ecuación; si lo están, entonces las covariables no están aportando al modelo todo lo que podrían. La independencia entre errores y covariables puede valorarse mediante diagramas de dispersión con cada covariable en el eje horizontal y los residuos en el vertical. Cuando los residuos son independientes de la covariable elegida, los puntos del diagrama están aleatoriamente repartidos en torno al valor cero del eje horizontal. Esto es lo que ocurre, por ejemplo, en el diagrama de dispersión representado en la Figura 5.7 correspondiente a la covariable *años consumiendo*.

Figura 5.7. Diagrama de dispersión: *años consumiendo* por *residuos*



Dispersión proporcional a la media

En regresión logística, cada observación Y se interpreta como un ensayo de Bernoulli con valor esperado π_1 y varianza $\pi_1(1 - \pi_1)$. Si el número de casos (n) es mayor que el número de patrones de variabilidad distintos (H), el número de eventos dentro de cada patrón de variabilidad se asume que se distribuye según el modelo de probabilidad binomial con índice n_h , valor esperado $n_h\pi_1$ y varianza $n_h\pi_1(1 - \pi_1)$.

Esto significa que en un análisis de regresión logística se está asumiendo que la varianza de cada patrón de variabilidad es proporcional a su media¹⁹, lo cual no es un problema cuando solo existe una observación por patrón de variabilidad (es decir, cuando el número de patrones de variabilidad es igual al número de casos), pero sí cuando a cada patrón de variabilidad le corresponde más de un caso (cosa que ocurre con datos agrupados, es decir, con covariables categóricas o con covariables cuantitativas que toman pocos valores). En estos casos es bastante habitual encontrar que la varianza observada no es proporcional a la media. Cuando la dispersión observada es mayor que la esperada decimos que existe **sobredispersión**; cuando es menor, **infradispersión**.

La dispersión observada y la esperada pueden ser distintas por diferentes motivos. Puede darse sobredispersión (la infradispersión es más bien poco frecuente) porque falta en el modelo alguna covariable importante, o porque hay subgrupos homogéneos de casos dentro de la muestra, es decir, observaciones no independientes entre sí, o porque la distribución de probabilidad elegida para el componente aleatorio no es apropiada para representar los datos, etc. (para profundizar en esta problemática recomendamos consultar Aitkin, Francis y Hinde, 2005; Gardner, Mulvey y Shaw, 1995; o McCullag y Nelder, 1989).

La sobredispersión es un problema porque hace que los errores típicos de las estimaciones sean más pequeños de lo que deberían, lo cual no solo altera la significación estadística de los valores estimados (aumenta el riesgo de declarar significativos efectos que no lo son) sino que hace que los intervalos de confianza de esos valores estimados sean más estrechos de lo que deberían (produciendo con ello una falsa impresión de precisión en las estimaciones).

El grado de dispersión suele cuantificarse mediante un parámetro de dispersión llamado *parámetro de escala*. Y este parámetro puede estimarse dividiendo la desvianza del modelo propuesto entre sus grados de libertad. Cuando la dispersión observada y la esperada son iguales, ese cociente toma un valor en torno a 1; un resultado mayor que 1 indica sobredispersión (valores mayores que 2 son problemáticos); un resultado menor que 1 indica infradispersión.

La desvianza y los grados de libertad necesarios para estimar el *parámetro de escala* pueden obtenerse con el procedimiento **Regresión logística multinomial** (tanto la desvianza como sus grados de libertad se ofrecen en la tabla *Estadísticos de bondad de ajuste*²⁰; el procedimiento **Regresión logística binaria** no ofrece esta desvianza). En nuestro ejemplo sobre la recuperación de pacientes con problemas de adicción, si se construye un modelo con las covariables *sexo*, *años* y *tto*, el valor que ofrece el procedimiento **Regresión logística multinomial** para la desvianza es 46,34, con 37 grados de libertad (para obtener esta información hay que marcar la opción **Bondad de ajuste** en el subcuadro de

¹⁹ Esta circunstancia contrasta con lo que ocurre en los modelos lineales clásicos. En el análisis de varianza o en el de regresión lineal, por ejemplo, se asume que la varianza de la variable dependiente es constante para cada patrón de variabilidad y, por tanto, independiente del valor de la media.

²⁰ La desvianza que se utiliza para estimar el parámetro de escala es la desvianza del modelo de regresión cuando se toma, como número de casos, el número de patrones de variabilidad distintos (datos agrupados), no cuando se considera que el número de patrones de variabilidad es el número de casos (datos no agrupados). El procedimiento **Regresión logística binaria** trabaja con datos no agrupados; el de **Regresión logística multinomial**, con datos agrupados.

diálogo *Regresión logística multinomial: Estadísticos*). Estos grados de libertad se obtienen restando 4 (el número de parámetros estimados) a los 41 patrones de variabilidad distintos que hay en el ejemplo. El cociente $46,34/37 = 1,25$ indica cierto grado de sobredispersión pero, puesto que es menor que 2, no parece que la sobredispersión sea un problema importante en los datos de nuestro ejemplo.

Los efectos indeseables de la sobredispersión pueden atenuarse aplicando una sencilla corrección a los errores típicos de los coeficientes. La corrección consiste en multiplicar cada error típico por la raíz cuadrada del valor estimado para el parámetro de escala (en nuestro ejemplo habría que multiplicar el error típico de cada coeficiente por la raíz cuadrada de 1,25). Esta corrección hace que los errores típicos sean ligeramente más grandes y, con ello, que aumente la amplitud de los intervalos de confianza y disminuya el riesgo de declarar significativos efectos que no lo son.

El procedimiento **Regresión logística multinomial** (se describe en el siguiente capítulo) ofrece la posibilidad de corregir automáticamente la dispersión observada aplicando bien una estimación del parámetro de escala basada en los datos, bien un valor concreto fijado por el usuario (ambas opciones están disponibles en el menú desplegable **Escala** del subcuadro de diálogo *Regresión logística multinomial: Opciones*).

Casos atípicos e influyentes

Valorar la calidad de una ecuación de regresión y, si fuera posible, mejorarla, requiere no solo vigilar el cumplimiento de los supuestos en los que se basa, sino controlar algunos detalles que podrían estar distorsionando los resultados del análisis, en concreto, la presencia de casos *mal pronosticados* y de casos *atípicos e influyentes*.

En el Capítulo 10 del segundo volumen (apartado *Casos atípicos e influyentes*) hemos presentado una serie de estadísticos para detectar la posible presencia de casos atípicos e influyentes en el contexto de la regresión lineal. Varios de estos estadísticos (distancias, medidas de influencia, etc.) han sido generalizados al ámbito de la regresión logística en un trabajo ya clásico de Pregibon (1981). No obstante, las peculiaridades de los modelos de regresión logística hacen que esta generalización no sea del todo satisfactoria. Consecuentemente, la interpretación de estos estadísticos debe realizarse con cautela (Fox, 1997; Hosmer y Lemeshow, 2000; Menard, 2001).

Casos atípicos

Al igual que en regresión lineal, también en regresión logística puede haber casos atípicos en la variable dependiente, en la(s) covariable(s) o en ambas.

El hecho de que la variable dependiente sea una variable dicotómica podría hacer pensar que no es posible encontrar valores atípicos en Y (pues todos los valores en Y son ceros y unos). Sin embargo, puede considerarse que un caso es atípico en Y cuando su valor, sea cero o uno, no se corresponde con lo que cabría esperar de él en función de

los valores que toma en el conjunto de las covariables X_j . En consecuencia, detectar casos atípicos en Y pasa por detectar *casos mal pronosticados*. Y éstos pueden detectarse revisando los **residuos** ($\hat{E}_i$), es decir, las diferencias entre las *probabilidades observadas*²¹ y las *probabilidades pronosticadas* por el modelo:

$$\hat{E}_i = P(Y) - \hat{\pi}_1 \quad [5.18]$$

(tanto la probabilidad observada, $P(Y)$, como la pronosticada, $\hat{\pi}_1$, se refieren a la categoría de referencia de la variable dependiente). Puesto que los residuos en bruto no son fácilmente interpretables, lo habitual es aplicarles algún tipo de transformación. Una de las más utilizadas consiste en dividirlos por su error típico. Se obtienen así los **residuos tipificados** o **estandarizados** ($\hat{E}_{Z_i}$), también llamados **residuos de Pearson** ($ZRE_ \#$ en SPSS):

$$\hat{E}_{Z_i} = \frac{\hat{E}_i}{\sqrt{\hat{\pi}_1 (1 - \hat{\pi}_1)}} \quad [5.19]$$

La distribución de estos residuos se aproxima a la normal tipificada tanto más cuanto mayor es el tamaño muestral. Por tanto, con muestras grandes cabe esperar que el 95% de estos residuos se encuentre entre -2 y 2 ; y el 99% entre $-2,5$ y $2,5$. Los residuos tipificados mayores que 3 o menores -3 corresponden a casos mal pronosticados. Y un caso mal pronosticado puede estar delatando la presencia de un caso atípico en Y .

Otros residuos muy utilizados en regresión logística son los **residuos de desviación** ($DEV_ \#$ en SPSS):

$$\hat{E}_{D_i} = \sqrt{-2 \log_e(\hat{\pi}_{\text{real}})} \quad [5.20]$$

(con los casos que pertenecen a la categoría codificada con un 1 se toma la raíz cuadrada *positiva*; con los que pertenecen a la categoría codificada con un 0 se toma la raíz cuadrada *negativa*). $\hat{\pi}_{\text{real}}$ se refiere a la probabilidad estimada de que un caso pertenezca a su grupo real, es decir, a la categoría de la variable dependiente a la que realmente pertenece²².

Los residuos de desviación son componentes de la desviación del modelo (sumándolos después de elevarlos al cuadrado se obtiene la desviación del modelo). Con muestras grandes, su distribución se aproxima a la distribución normal tipificada; por tanto, pueden interpretarse exactamente igual que los residuos tipificados.

Aunque ambos tipos de residuos se parecen, hay dos razones para preferir los de *desviación* a los *tipificados*. En primer lugar, la distribución de los residuos de desviación se parece a la distribución normal más de lo que se parece la distribución de los residuos

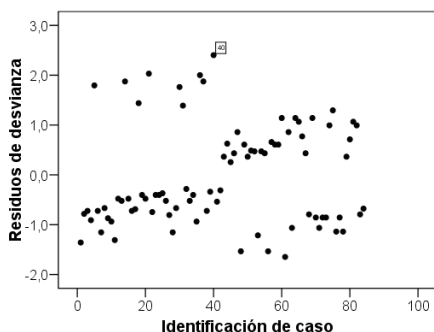
²¹ En una regresión logística binaria con datos no agrupados, la probabilidad observada siempre vale 1 para los casos que pertenecen a la categoría de referencia y 0 para los restantes casos.

²² Esta probabilidad puede obtenerse, si se tuviera interés en ella, marcando la opción **Probabilidad de la categoría real** en el subcuadro de diálogo *Regresión logística multinomial: Guardar*.

tipificados. En segundo lugar, cuando las probabilidades pronosticadas se encuentran cerca de cero o uno, los residuos tipificados son algo inestables.

En el diagrama de dispersión de la Figura 5.8 están representados los residuos de *desvianza* de nuestro ejemplo²³ (con los residuos *tipificados* se obtiene una nube de puntos muy parecida). El diagrama muestra que no existen residuos menores que -2 y que ninguno de ellos es mayor que 2,5. Por tanto, no parece que haya casos especialmente mal pronosticados y, consecuentemente, no parece que existan casos atípicos en Y . El caso identificado en el gráfico (el caso nº 40) es al que corresponde el residuo de desvianza más alejado de cero (2,40).

Figura 5.8. Diagrama de dispersión con los residuos de desvianza



Para detectar casos inusuales o atípicos en las covariables puede utilizarse, al igual que en regresión lineal, un estadístico llamado **influencia** (*leverage*; $LEV_ \#$ en SPSS). Este estadístico refleja el grado de alejamiento de cada caso respecto del centro de su distribución en el conjunto de covariables.

Los valores de influencia de una regresión logística oscilan entre 0 y 1, y su media vale $(p+1)/n$ (donde p se refiere al número de covariables). Cuanto más alejado se encuentra un caso del centro de su distribución, mayor es su valor de influencia²⁴ y, consecuentemente, más inusual o atípico es en X_j .

Para interpretar el tamaño de los valores de influencia puede servir de guía lo ya dicho a propósito de la regresión lineal (ver Capítulo 10 del segundo volumen). Stevens

²³ Estos residuos se obtienen marcando la opción **Desvianza** del subcuadro de diálogo *Regresión logística binaria: Guardar*.

²⁴ En regresión lineal, cuanto mayor es el valor de influencia de un caso, más alejado se encuentra del centro de su distribución. En regresión logística no ocurre exactamente esto. El valor de influencia de un caso no viene determinado únicamente por las variables independientes, sino también por la dependiente. Y esto tiene sus consecuencias. En regresión logística, el valor de influencia de un caso es tanto mayor cuanto más alejado se encuentra ese caso del centro de su distribución, pero hasta un punto a partir del cual el valor de influencia disminuye rápidamente. Esto significa que casos extremadamente alejados del centro de su distribución pueden tener valores de influencia más pequeños que casos no tan alejados. Por tanto, para interpretar el valor de influencia de un caso hay que prestar atención a su probabilidad pronosticada: únicamente de los casos con probabilidades pronosticadas comprendidas entre 0,10 y 0,90 puede asegurarse que el valor de influencia está reflejando su alejamiento del resto de los casos.

(1992) sugiere revisar los casos con valores de influencia mayores que $3(p+1)/n$. Y una regla que funciona razonablemente bien para identificar casos atípicos en X_j es la siguiente: los valores menores que 0,2 son poco problemáticos, los valores comprendidos entre 0,2 y 0,5 son arriesgados; los valores mayores que 0,5 deben revisarse.

En nuestro ejemplo, hay un caso (el nº 1) cuyo valor de influencia es 0,19; los valores de influencia del resto de los casos no llegan a 0,10. Por tanto, no parece que haya que preocuparse por la presencia de casos atípicos en las covariables.

Casos influyentes

Determinar la influencia de un caso en la ecuación de regresión pasa por comparar los resultados que se obtienen con la ecuación que incluye todos los casos con los resultados que se obtienen al ir eliminando cada caso de la ecuación (en caso necesario, revisar el concepto de *influencia* en el apartado *Casos influyentes* del Capítulo 10 del segundo volumen).

Una buena forma de obtener alguna evidencia sobre la influencia de cada caso consiste en valorar el cambio que se produce en el **ajuste global del modelo** al ir eliminando casos. Este cambio puede cuantificarse comparando la desviación del modelo propuesto ($-2LL_1$) con esa misma desviación al eliminar cada caso del análisis ($-2LL_{1(i)}$). La diferencia entre estas dos desviaciones será tanto mayor cuanto mayor sea la contribución de un caso al ajuste del modelo. Y esta diferencia puede estimarse a partir de los **residuos studentizados** ($SRE_#$ en SPSS):

$$\hat{E}_{S_i} = \frac{\hat{E}_i}{\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)(1-h_i)}} \quad [5.21]$$

(h_i se refiere a los valores de influencia). Estos residuos, elevados al cuadrado, son una buena estimación del cambio que se produce en la desviación al ir eliminando casos. Con muestras grandes se distribuyen de forma aproximadamente normal. Por tanto, residuos studentizados mayores que 3 en valor absoluto suelen estar delatando, por lo general, casos excesivamente influyentes.

Otra forma de valorar la influencia de un caso en la ecuación de regresión consiste en cuantificar cómo afecta su ausencia al tamaño de los coeficientes. El cambio en los coeficientes puede valorarse de forma individual o de forma colectiva. La influencia de un caso sobre cada coeficiente de regresión puede valorarse a partir de la **diferencia entre los coeficientes de regresión** ($DFB_#$ en SPSS). Y el cambio que experimentan todos los coeficientes de regresión de forma simultánea o conjunta puede valorarse con una medida análoga a la **distancia de Cook** ($COO_#$ en SPSS):

$$D_{Cook} = (\hat{E}_{Z_i}^2 \times h_i) / (1 - h_i)$$

Los casos con una distancia de Cook mayor que 1 deben ser revisados (es probable que se trate de casos influyentes). En nuestro ejemplo, ningún residuo *studentizado* es menor

que -2 , solo tres son mayores que 2 y ninguno es mayor que 3. Cuatro casos tienen distancias de Cook mayores que 0,20 (entre ellos, el caso n° 40; ver Figura 5.8), pero ninguna distancia es mayor que 0,50. Por tanto, no parece que en nuestro ejemplo haya casos excesivamente influyentes.

Apéndice 5

Regresión probit

Ya hemos argumentado al principio del capítulo que una ecuación lineal no es una estrategia adecuada para modelar respuestas dicotómicas. Se obtienen mejores resultados con ecuaciones que, al definir una relación curvilínea, ofrecen pronósticos comprendidos dentro del rango 0-1. Entre estas ecuaciones, la *función logística* es la más utilizada, pero no es la única. Cualquier función de probabilidad acumulada monótona creciente ofrece valores dentro del rango 0-1. Y, entre éstas, la **función probit** es la que ha recibido más atención.

La función *probit* modela $P(Y=1)$ o, más brevemente, π_1 , a partir de las probabilidades acumuladas correspondientes a cada pronóstico lineal:

$$P(Y=1) = \pi_1 = F(\beta_0 + \beta_1 X) \quad [5.22]$$

La peculiaridad de esta ecuación es que F se refiere a las probabilidades acumuladas de una *distribución normal*. La curva de regresión que se obtiene con [5.22] tiene la forma de una función de densidad de probabilidad acumulada; por tanto, se parece bastante a la curva que se obtiene con una ecuación logística.

La ecuación [5.22] se vuelve lineal al modelar la función inversa de π_1 . Precisamente la forma inversa de esa ecuación es la expresión habitual de la función *probit*:

$$probit(Y=1) = F^{-1}(\pi_1) = \beta_0 + \beta_1 X \quad [5.23]$$

Esta ecuación devuelve la puntuación Z que acumula, en una curva normal tipificada, una proporción de casos (área bajo la curva) igual a π_1 . Por ejemplo, en una curva normal tipificada, la puntuación $Z=0$ acumula una proporción de casos de 0,50; por tanto, $probit(0,50) = 0$. La puntuación $Z=1,64$ acumula una proporción de casos de 0,95; por tanto, $probit(0,95) = 1,64$. Etc.

Tanto $P(Y=1)$ como $logit(Y=1)$ y $probit(Y=1)$ están expresando la misma idea, pero en distinta escala. Esto puede apreciarse en los valores que ofrece la Tabla 5.34. Una probabilidad toma valores comprendidos entre cero y uno, y cada valor es simétrico de su complementario (a una probabilidad de 0,25 le corresponde un valor complementario de $1 - 0,25 = 0,75$). Un *logit* no tiene ni mínimo ni máximo (en teoría, toma valores entre $-\infty$ y $+\infty$); a una probabilidad de 0,50 le corresponde un *logit* de 0; y los valores son simétricos respecto de 0. Un *probit* se comporta de forma muy parecida a un *logit*: no tiene mínimo ni máximo, a una probabilidad de 0,50 le corresponde un *probit* de 0 y los valores son simétricos respecto de 0.

Tabla 5.34. Relación entre *probabilidad*, *logit* y *probit*

<i>Prob</i> ($Y = 1$)	<i>Logit</i> ($Y = 1$)	<i>Probit</i> ($Y = 1$)
0,01	-4,60	-2,33
0,10	-2,20	-1,28
0,25	-1,10	-0,67
0,50	0,00	0,00
0,75	1,10	0,67
0,90	2,20	1,28
0,99	4,60	2,33

Las funciones *logit* y *probit* ofrecen resultados (pronósticos y ajuste) muy parecidos. Pero en igualdad de condiciones, los valores de los coeficientes de regresión son más pequeños en el caso de la función *probit* que en el de la función *logit*. Esto es debido a que la distribución logística es más dispersa que la distribución normal (esto también se aprecia en los datos de la Tabla 5.34). Ambas distribuciones tienen media 0, pero la desviación típica vale 1 en el caso de la distribución normal tipificada y 1,8 en el de la distribución logística. Cuando ambas funciones se ajustan bien a los datos, el tamaño de las estimaciones de una ecuación logística es aproximadamente 1,8 veces mayor que las de una ecuación *probit*.

El SPSS incluye varios procedimientos para ajustar modelos de regresión *probit*. La opción **Regresión > Probit** (procedimiento PROBIT) requiere que los datos estén agrupados y no guarda las probabilidades pronosticadas (las ofrece en una tabla de resultados). Las opciones **Regresión > Ordinal** (procedimiento PLUM) y **Modelos lineales generalizados** (procedimiento GENLIN) permiten ajustar modelos de regresión *probit* con datos agrupados y no agrupados, y guardar las probabilidades pronosticadas en una variable del archivo de datos (en ambos casos es necesario elegir explícitamente *probit* como función de enlace pues, en estos dos procedimientos, no es la función de enlace que se aplica por defecto).

Retomemos nuestro ejemplo sobre 84 pacientes con problemas de adicción al alcohol (archivo *Tratamiento adicción alcohol*). Al ajustar un modelo de regresión logística con *recuperación* como variable dependiente y *tto* (tratamiento) como covariable hemos obtenido la siguiente ecuación de regresión (ver Tabla 5.12):

$$\text{logit}(\text{recuperación} = 1) = -1,30 + 1,89(\text{tto})$$

Al ajustar un modelo de regresión *probit* a los mismos datos se obtiene una ecuación bastante parecida:

$$\text{probit}(\text{recuperación} = 1) = -0,79 + 1,16(\text{tto}) \quad [5.24]$$

(los coeficientes de regresión son significativamente distintos de cero tanto en la ecuación logística como en la *probit*). Ya sabemos que los coeficientes de una ecuación logística se interpretan transformándolos en *odds ratios*. Los coeficientes de una ecuación *probit* se interpretan transformándolos en probabilidades. Así, con el tratamiento *estándar* ($\text{tto} = 0$), la ecuación *probit* ofrece un pronóstico de -0,79. La probabilidad acumulada hasta la puntuación -0,79 en una curva normal tipificada vale 0,21. Por tanto, la ecuación [5.24] estima que la probabilidad de *recuperación* con el tratamiento *estándar* vale 0,21. Esta probabilidad de recuperación con el tratamiento *estándar* es idéntica a la estimada con la ecuación logística (ver ecuación [5.12]).

El pronóstico que ofrece la ecuación [5.24] para el tratamiento *combinado* ($tto = 1$) vale $-0,79 + 1,16 = 0,36$. La probabilidad acumulada hasta la puntuación 0,36 en una curva normal tipificada vale 0,64. Por tanto, la ecuación [5.24] estima que la probabilidad de *recuperación* con el tratamiento *combinado* vale 0,64. Y esta probabilidad también es idéntica a la estimada con la ecuación logística (ver ecuación [5.12]).

Al incluir más de una variable independiente en la ecuación se mantiene el parecido entre ambas ecuaciones. Cuando hemos ajustado un modelo de regresión logística con la *recuperación* como variable dependiente y el *sexo*, los años consumiendo (*años*) y el tratamiento (*tto*) como covariables, hemos obtenido la siguiente ecuación de regresión (ver Tabla 5.20):

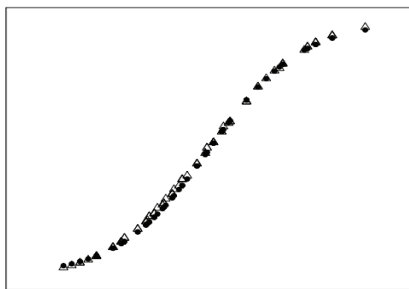
$$\text{logit}(\text{recuperación} = 1) = 2,11 - 1,33(\text{sexo}) - 0,18(\text{años}) + 1,84(\text{tto}) \quad [5.25]$$

Al ajustar un modelo de regresión *probit* a los mismos datos se obtiene una ecuación bastante parecida:

$$\text{probit}(\text{recuperación} = 1) = 1,27 - 0,80(\text{sexo}) - 0,11(\text{años}) + 1,27(\text{tto}) \quad [5.26]$$

(los coeficientes de regresión son significativamente distintos de cero tanto en la ecuación logística como en la *probit*). El parecido entre ambas ecuaciones es evidente, sobre todo si se tiene en cuenta que la dispersión de una distribución logística es 1,8 veces mayor que la de una distribución normal. Y cuando los pronósticos *logit* de [5.25] y los pronósticos *probit* de [5.26] se transforman en sus correspondientes probabilidades, es difícil, tal como muestra la Figura 5.9, distinguir unas de otras.

Figura 5.9. Relación entre cada patrón de variabilidad (eje horizontal) y las probabilidades pronosticadas por un modelo *logit* (círculos negros) y un modelo *probit* (triángulos blancos)



Regresión logística (II).

Respuestas nominales y ordinales

Acabamos de ver que la regresión logística **binaria** o **dicotómica** sirve para modelar respuestas *dicotómicas*. Para modelar respuestas *politómicas* (variables categóricas con más de dos categorías) suele utilizarse una extensión de la regresión logística binaria llamada regresión logística **nominal**, **politómica** o **multinomial** (ver McFaden, 1974; Agresti, 2002, 2007). Y si las categorías de la variable están cuantitativamente ordenadas, entonces puede utilizarse otra versión de la regresión logística llamada **regresión ordinal** (ver Agresti, 2010; Clogg y Shihadeh, 1994; Long, 1997).

Regresión nominal

Ya sabemos que el *análisis de regresión logística* sirve para pronosticar los valores de una variable dependiente *categórica* a partir de una o más variables independientes *categóricas* o *cuantitativas*. Hemos visto que, con variables dependientes dicotómicas, la regresión logística viene acompañada de los calificativos *binaria* o *dicotómica*. Cuando la variable dependiente es politómica (categórica con más de dos categorías), el correspondiente análisis de regresión logística recibe el nombre de **nominal**, **politómica** o **multinomial**.

Con *nominal* se está poniendo el énfasis en el nivel de medida de la variable dependiente; con *politómica* se está destacando el hecho de que la variable dependiente tiene más de dos categorías (lo cual sirve para distinguir esta versión de la estudiada en el capítulo anterior); con *multinomial* se está haciendo referencia a uno de los supuestos

básicos del análisis: en cada patrón de variabilidad (en cada combinación distinta entre variables independientes), las frecuencias de las categorías de la variable dependiente se asume que se distribuyen según el modelo de probabilidad multinomial.

El modelo de regresión nominal

Entender cómo funciona la regresión logística nominal es una tarea relativamente sencilla cuando ya se sabe cómo funciona la regresión logística binaria: la versión nominal no es más que una sucesión de $K - 1$ versiones binarias, siendo K el número de categorías de la respuesta que se desea modelar.

Seguimos, por tanto, trabajando con un modelo lineal generalizado con función de enlace logit, pero con una importante diferencia respecto de lo que ya conocemos: en lugar de utilizar una sola ecuación para modelar la comparación entre las dos categorías de una respuesta dicotómica, se utilizan $K - 1$ ecuaciones para modelar la comparación entre las K categorías de una respuesta politómica; pero no la comparación de cada categoría con cada otra, sino de cada una con otra, siempre la misma (generalmente la primera o la última), que se toma como categoría de referencia.

Siendo π_k la probabilidad teórica asociada a cada categoría de la variable dependiente y tomando la última categoría, K , como categoría de referencia, pueden definirse $K - 1$ funciones logit no redundantes del tipo¹:

$$\log_e \left[\frac{\pi_k}{\pi_K} \right] = \beta_{0k} + \beta_{1k}X_1 + \beta_{2k}X_2 + \cdots + \beta_{jk}X_j + \cdots + \beta_{pk}X_p \quad [6.1]$$

donde K se refiere a la última categoría de la variable y k a cualquiera de las restantes. Los coeficientes de regresión aparecen con dos subíndices porque en [6.1] se está definiendo más de una ecuación logit. El primer subíndice ($j = 1, 2, \dots, p$) sirve para identificar cada una de las p variables independientes; el segundo ($k = 1, 2, \dots, K - 1$), para identificar cada una de las $K - 1$ ecuaciones logit.

Una variable independiente (regresión simple)

En nuestro ejemplo sobre 84 pacientes con problemas de adicción (archivo *Tratamiento adicción alcohol*) hay una variable politómica llamada *recaída*. Esta variable tiene tres categorías que sirven para identificar a los pacientes que, una vez finalizado el tratamiento, han recaído durante el primer año (código 1), han recaído durante el segundo año (código 2) y no han recaído en los dos primeros años (código 3). La Tabla 6.1 muestra un resumen de esta variable combinada con la variable *tto* (tratamiento).

¹ Cuando la variable dependiente es dicotómica basta con utilizar una ecuación de regresión, pues intercambiando la categoría de referencia se obtiene exactamente la misma ecuación con los coeficientes cambiados de signo. Cuando la variable dependiente tiene K categorías, hay $K - 1$ ecuaciones con información no redundante (la K -ésima ecuación no aporta información nueva). Cuando $K = 2$, la ecuación [6.1] equivale al modelo de regresión logística binaria.

Los porcentajes de fila indican que, de los 42 pacientes que han recibido el tratamiento estándar, dos tercios recaen a lo largo del primer año y solamente el 11,9% no recaen; y de los 42 pacientes que han recibido el tratamiento combinado, un tercio recaen a lo largo del primer año y algo más de la mitad, el 52%, no recaen.

El estadístico ji-cuadrado de Pearson aplicado a estos datos permite rechazar la hipótesis de independencia entre *tto* y *recaída* ($p < 0,0005$); y esto significa que los porcentajes de las categorías de *recaída* no son iguales con ambos tratamientos. Un modelo de regresión logística puede aclarar en qué sentido no son iguales.

Tabla 6.1. Frecuencias conjuntas de *tratamiento* por *recaída*

			Recaída			Total
			Primer año	Segundo año	No recaen	
Tratamiento	Estándar	Recuento	28	9	5	42
		% de Tratamiento	66,7%	21,4%	11,9%	100,0%
	Combinado	Recuento	14	6	22	42
		% de Tratamiento	33,3%	14,3%	52,4%	100,0%
Total		Recuento	42	15	27	84
		% de Tratamiento	50,0%	17,9%	32,1%	100,0%

Dado que la variable *recaída* tiene $K = 3$ categorías, para analizarla mediante un modelo de regresión logística es necesario formular $K - 1 = 2$ ecuaciones. Podemos llamar a estas ecuaciones, para distinguirlas, logit_1 y logit_2 . Tomando la última categoría (no recaen) como categoría de referencia,

$$\begin{aligned}\text{logit}_1 &= \log_e(\pi_{\text{primer año}} / \pi_{\text{no recaen}}) = \beta_{01} + \beta_{11}(\text{tto}) \\ \text{logit}_2 &= \log_e(\pi_{\text{segundo año}} / \pi_{\text{no recaen}}) = \beta_{02} + \beta_{12}(\text{tto})\end{aligned}\quad [6.2]$$

En ambas ecuaciones se está modelando cómo cambia el *logit de recaer* a partir del tratamiento recibido. Pero en el primer caso se está modelando el logit de recaer el *primer año* y en el segundo caso el logit de recaer el *segundo año* (en ambos casos las *odds* del interior del paréntesis se calculan respecto de la categoría *no recaer*).

Para ajustar con el SPSS un modelo de regresión logística multinomial con *recaída* como variable dependiente y *tto* como variable independiente:

- Seleccionar la opción **Regresión > Logística multinomial** del menú **Analizar** para acceder al cuadro de diálogo *Regresión logística multinomial*.
- Trasladar la variable *recaída* al cuadro **Dependiente** (dejar como categoría de referencia la que el programa asigna por defecto, es decir, la última) y la variable *tto* a la lista **Factores**².

² Puesto que la variable *tto* es dicotómica, puede incluirse indistintamente como *factor* o como *covariable*. De ambas formas se obtiene el mismo resultado, pero hay que vigilar, en la interpretación, cuál es la categoría de referencia (pues la *odds ratio* puede calcularse tanto dividiendo *estándar* entre *combinado* como *combinado* entre *estándar*).

Aceptando estas selecciones se obtienen los resultados que muestran las Tablas 6.2 a 6.5. La Tabla 6.2 ofrece un resumen (frecuencias absolutas y porcentuales) de las variables incluidas en el análisis (*recaída y tratamiento*) y el número de patrones de variabilidad (*subpoblaciones*), que con una variable independiente dicotómica son solo 2.

Tabla 6.2. Resumen de los casos procesados

		N	% marginal
Recaída	Primer año	42	50,0%
	Segundo año	15	17,9%
	No recae	27	32,1%
Tratamiento	Estándar	42	50,0%
	Combinado	42	50,0%
Válidos		84	100,0%
Perdidos		0	
Total		84	
Subpoblación		2	

Ajuste global

La Tabla 6.3 contiene la información necesaria para realizar una valoración global del modelo, es decir, para decidir si el conjunto de variables independientes incluidas en el análisis (de momento, solo *tto*) contribuyen o no a reducir el desajuste del modelo nulo. La tabla incluye la desviación del modelo nulo (*sólo la intersección*: $-2LL_0 = 31,55$), la desviación del modelo propuesto (*final*: $-2LL_1 = 14,63$) y la diferencia entre ambas, es decir, la razón de verosimilitudes G^2 (*chi-cuadrado*; ver ecuación [5.9]):

$$G^2 = -2LL_0 - (-2LL_1) = 31,55 - 14,63 = 16,92$$

Este estadístico permite contrastar la hipótesis nula de que los términos en que difieren el modelo nulo y el modelo propuesto valen cero en la población. El rechazo de esta hipótesis estaría indicando que el modelo propuesto contribuye a reducir el desajuste del modelo nulo. En nuestro ejemplo, el nivel crítico asociado a la razón de verosimilitudes (*sig.* < 0,0005) permite rechazar la hipótesis de que el coeficiente de regresión asociado a la variable *tto* vale cero en la población y, consecuentemente, se puede concluir que la variable *tto* contribuye a reducir el desajuste del modelo nulo.

Tabla 6.3. Estadísticos de ajuste global: desviación y razón de verosimilitudes

Modelo	Criterio de ajuste del modelo	Contrastes de la razón de verosimilitud		
	-2 log verosimilitud	Chi-cuadrado	gl	Sig.
Sólo la intersección	31,55			
Final	14,63	16,92	2	,000

Los estadísticos tipo R^2 que ofrece la Tabla 6.4 permiten cuantificar en qué medida se consigue reducir el desajuste del modelo nulo. El estadístico de Nagelkerke indica que la variable *tto* consigue reducir ese desajuste en un 21%.

Tabla 6.4. Estadísticos de ajuste global: pseudo *R*-cuadrado

Cox y Snell	,18
Nagelkerke	,21
McFadden	,10

Significación e interpretación de los coeficientes de regresión

La última de las tablas que ofrece el *Visor* contiene las estimaciones de los coeficientes del modelo (ver Tabla 6.5). Puesto que la variable dependiente *recaída* tiene tres categorías, la tabla ofrece dos ecuaciones de regresión (ver [6.2]). Al definir la variable *tto* como *factor*, el programa fija en cero la categoría con el código mayor (*tto combinado*) y únicamente estima el coeficiente de la otra categoría (*tto estándar*):

$$\begin{aligned}\text{logit}_1 &= \log_e(\hat{\pi}_{\text{primer año}} / \hat{\pi}_{\text{no recaer}}) = -0,45 + 2,17 (\text{estándar}) \\ \text{logit}_2 &= \log_e(\hat{\pi}_{\text{segundo año}} / \hat{\pi}_{\text{no recaer}}) = -1,30 + 1,89 (\text{estándar})\end{aligned}\quad [6.3]$$

Tanto el estadístico de Wald como los correspondientes intervalos de confianza indican que el coeficiente asociado al tratamiento estándar (identificado como *tto* = 0) es significativamente distinto de cero (*sig.* < 0,0005 en el primer logit y *sig.* = 0,009 en el segundo).

Tabla 6.5. Estimaciones de los parámetros

Recaída ^a		B	Error típ.	Wald	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B)	
								L. inferior	L. superior
Primer año	Intersección	-,45	,34	1,75	1	,186			
	[tto=0]	2,17	,59	13,41	1	,000	8,80	2,75	28,18
	[tto=1]	0 ^b	.	.	0	.	.	.	.
Segundo año	Intersección	-1,30	,46	7,96	1	,005			
	[tto=0]	1,89	,72	6,81	1	,009	6,60	1,60	27,24
	[tto=1]	0 ^b	.	.	0	.	.	.	.

a. La categoría de referencia es: No recaer.

b. Este parámetro se ha establecido a cero porque es redundante.

La intersección es, en ambos casos, el logit estimado para el tratamiento combinado, es decir, el logit estimado para la categoría cuyo coeficiente de regresión se ha fijado en cero (que en nuestro ejemplo es *tto* = 1]. El signo negativo de las intersecciones está indicando que, con el tratamiento combinado, recaer en el primer año (categoría 1) y en el segundo (categoría 2) es menos probable que no recaer (categoría de referencia).

Los valores exponenciales de las intersecciones ($e^{-0,45} = 0,64$ en el primer logit y $e^{-1,30} = 0,27$ en el segundo; estos valores no los ofrece la tabla) indican que, entre los pacientes que reciben el tratamiento combinado, la proporción de recaídas durante el primer año es un 64% de la proporción de no recaídas, y la proporción de recaídas durante el segundo año es un 27% de la proporción de no recaídas.

Toda esta información puede obtenerse fácilmente a partir de las frecuencias de la Tabla 6.1. Por ejemplo, entre quienes reciben el tratamiento combinado, la *odds* de recaer el *primer año* respecto de *no recaer* vale $14/22 = 0,64$, que es el valor exponencial de la intersección del primer logit. Y el logaritmo de esta *odds* es $-0,45$, que es el valor estimado para la intersección en el primer logit.

Aunque la intersección contiene información interesante, no dice nada acerca de la relación entre la variable independiente (*tto*) y la dependiente (*recaída*). Para esto hay que fijarse en el coeficiente de regresión correspondiente a la categoría de la variable independiente que no se ha fijado en cero (en nuestro ejemplo, el tratamiento estándar: *tto* = 0). El signo positivo del coeficiente en el primer logit (2,17) indica que la *odds* de recaer el *primer año* aumenta con el tratamiento *estándar* (debe tenerse presente que la *odds* se calcula siempre respecto de la categoría *no recaer*, que es la categoría de referencia). Y el valor exponencial del coeficiente ($e^{2,17} = 8,80$) permite concretar que la *odds* de recaer el *primer año* es 8,80 veces mayor con el tratamiento estándar que con el combinado.

En el segundo logit está ocurriendo algo parecido. El signo positivo del coeficiente (1,89) indica que la *odds* de recaer el *segundo año* aumenta con el tratamiento estándar. Y el valor exponencial del coeficiente ($e^{1,89} = 6,60$) permite concretar que la *odds* de recaer el *segundo año* es 6,60 veces mayor con el tratamiento estándar que con el combinado³.

Las dos ecuaciones de la Tabla 6.5 describen la relación entre las variables *tto* y *recaída* comparando las dos primeras categorías de la variable dependiente con la tercera (es decir, comparando las recaídas en el *primer* y *segundo año* con las *no recaídas*). En el caso de que interese realizar la comparación que falta (las dos primeras categorías entre sí, es decir, las recaídas en el *primer* y *segundo año*), puede repetirse el análisis cambiando la categoría de referencia de la variable dependiente.

Eliendo como categoría de referencia la primera (recaer el *primer año*), el coeficiente asociado al logit que compara la segunda categoría con la primera (recaer el *segundo año* respecto de recaer el *primer año*) vale $-0,29$. El signo negativo del coeficiente indica que la *odds* de recaer el *segundo año* (ahora, respecto de recaer el *primer*

³ Una vez más conviene recordar que no debe confundirse el cambio en las *odds* con el cambio en las *probabilidades* (los cálculos que se ofrecen a continuación se basan en las frecuencias de la Tabla 6.1). La *odds* de recaer el *primer año* respecto de *no recaer* vale $28/5 = 5,60$ cuando se recibe el tratamiento *estándar* y $14/22 = 0,636$ cuando se recibe el *combinado*; de ahí que el análisis de regresión logística esté indicando que una *odds* es 8,80 veces mayor que la otra ($5,60/0,636 = 8,80$). Del mismo modo, la *odds* de recaer el *segundo año* respecto de *no recaer* vale $9/5 = 1,80$ cuando se recibe el tratamiento *estándar* y $6/22 = 0,273$ cuando se recibe el *combinado*; de ahí que el análisis de regresión logística esté indicando que una *odds* es 6,60 veces mayor que la otra ($1,80/0,273 = 6,60$). Sin embargo, la probabilidad de recaer el *primer año* vale $28/42 = 0,667$ con el tratamiento *estándar* y $14/42 = 0,333$ con el *combinado*, es decir, solamente el doble, no 8,80 veces más. Y la probabilidad de recaer el *segundo año* vale $9/42 = 0,214$ con el tratamiento *estándar* y $6/42 = 0,143$ con el *combinado*, es decir, solamente 1,5 veces más, no 6,60 veces más.

año, no respecto de *no recaer*) disminuye con el tratamiento estándar. Y el valor exponencial del coeficiente ($e^{-0.29} = 0,75$) indica que la *odds* de recaer en el *segundo año* con el tratamiento estándar es un 75 % de esa misma *odds* con el tratamiento combinado. No obstante, esta diferencia no alcanza la significación estadística (*sig.* = 0,643); por tanto, no existe evidencia de que la proporción de recaídas en el *segundo año* respecto del *primer año* cambie por aplicar uno u otro tratamiento.

Más de una variable independiente (regresión múltiple)

Veamos ahora cómo ajustar e interpretar un modelo de regresión nominal múltiple añadiendo al modelo propuesto en el apartado anterior (un modelo que únicamente incluía la covariable *tto*) las variables *sexo* y *años* (años consumiendo). De nuevo, puesto que la variable *recaída* tiene $K = 3$ categorías, para poder modelarla mediante una regresión logística es necesario formular $K - 1 = 2$ ecuaciones:

$$\begin{aligned}\text{logit}_1 &= \log_e \left[\frac{\pi_{\text{primer año}}}{\pi_{\text{no recaer}}} \right] = \beta_{01} + \beta_{11}(\text{tto}) + \beta_{21}(\text{sexo}) + \beta_{31}(\text{años}) \\ \text{logit}_2 &= \log_e \left[\frac{\pi_{\text{segundo año}}}{\pi_{\text{no recaer}}} \right] = \beta_{02} + \beta_{12}(\text{tto}) + \beta_{22}(\text{sexo}) + \beta_{32}(\text{años})\end{aligned}\quad [6.4]$$

En ambas ecuaciones se está modelando cómo cambia el *logit de recaer* a partir del *tratamiento* recibido, del *sexo* y del número de *años de consumo*. Pero en el primer caso se está modelando el logit de recaer el *primer año* y en el segundo caso el logit de recaer el *segundo año* (las *odds* del interior del paréntesis se calculan, en ambos casos, respecto de la categoría *no recaer*).

Para ajustar con el SPSS un modelo de regresión logística multinomial con *recaída* como variable dependiente y *tto*, *sexo* y *años* como variables independientes:

- Seleccionar la opción **Regresión > Logística multinomial** del menú **Analizar** para acceder al cuadro de diálogo *Regresión logística multinomial*.
- Trasladar la variable *recaída* al cuadro **Dependiente** (dejar como categoría de referencia la que el programa asigna por defecto, es decir, la última) y las variables *tto*, *sexo* y *años_c* (años consumiendo *centrada*) a la lista **Covariables**⁴.
- Pulsar el botón **Estadísticos** para acceder al subcuadro de diálogo *Regresión logística multinomial: Estadísticos* y marcar las opciones **Tabla de clasificación** y **Bondad de ajuste**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

⁴ Las variables independientes *categoricas* deben ser tratadas como *factores*; las *cuantitativas*, como *covariables*. Las variables *dicotómicas* pueden ser tratadas indistintamente como *factores* y como *covariables*. Ya hemos visto en el apartado anterior cómo se interpreta una variable dicotómica (*tto*) cuando se define como un *factor*; en este apartado vamos a ver cómo se interpreta cuando se define como una covariable. Hay detalles que cambian.

Aceptando estas selecciones se obtienen, entre otros, los resultados que muestran las Tablas 6.6 a 6.11. Exceptuando los estadísticos de bondad de ajuste de la Tabla 6.8 y los resultados de la de clasificación de la Tabla 6.11, el resto de la información ya se ha discutido en el apartado anterior a propósito del modelo de regresión simple.

Ajuste global

Los resultados que ofrece el SPSS incluyen tres tablas con información sobre el ajuste global del modelo. Las Tablas 6.6 y 6.7 se obtienen por defecto; la Tabla 6.8 se obtiene marcando la opción **Bondad de ajuste**.

Los estadísticos de la Tabla 6.6 sirven para decidir si el conjunto de variables independientes incluidas en el análisis (*tto*, *sexo* y *años*) contribuyen o no a reducir el desajuste del modelo nulo. La tabla ofrece la desviación del modelo nulo (*sólo la intersección*: $-2LL_0 = 142,19$), la desviación del modelo propuesto (*final*: $-2LL_1 = 95,21$) y la diferencia entre ambas desviaciones, es decir, la razón de verosimilitudes G^2 (*chi-cuadrado* = 46,98). El estadístico G^2 permite contrastar la hipótesis nula de que todos los coeficientes de regresión en que difieren el modelo nulo y el propuesto (es decir, todos los coeficientes de regresión del modelo propuesto, excluida la constante) valen cero en la población.

En el ejemplo, el nivel crítico asociado a la razón de verosimilitudes (*sig.* < 0,0005) permite rechazar la hipótesis nula de que todos los coeficientes de regresión valen cero. Puede concluirse, por tanto, que las variables independientes incluidas en la ecuación contribuyen a reducir el desajuste del modelo nulo. El valor del estadístico de Nagelkerke (Tabla 6.7) indica que esa reducción del desajuste alcanza el 49%.

Tabla 6.6. Estadísticos de ajuste global: desviación y razón de verosimilitudes

Modelo	Criterio de ajuste del modelo	Contrastes de la razón de verosimilitud		
	-2 log verosimilitud	Chi-cuadrado	gl	Sig.
Sólo la intersección	142,19			
Final	95,21	46,98	6	,000

Tabla 6.7. Estadísticos de ajuste global: pseudo *R*-cuadrado

Cox y Snell	,43
Nagelkerke	,49
McFadden	,27

Los estadísticos de la Tabla 6.8 permiten hacer una valoración del ajuste del modelo a partir de la comparación de los valores observados y los pronosticados. Esta forma de valorar el ajuste del modelo es complementaria de la que ofrece la Tabla 6.6. La razón de verosimilitudes de la Tabla 6.6 se obtiene comparando el modelo propuesto con el modelo nulo (el modelo que solo incluye la intersección); por tanto, permite valorar en

qué medida el modelo propuesto consigue reducir el desajuste del modelo nulo. Esto contrasta con los estadísticos de la Tabla 6.8, que se obtienen comparando el modelo propuesto con el modelo saturado⁵ (el modelo con ajuste perfecto) y, por tanto, sirven para valorar en qué medida el modelo propuesto se aleja del ajuste perfecto.

En nuestro ejemplo tenemos 41 patrones de variabilidad (es decir, 41 combinaciones distintas entre *tto*, *sexo* y *años_c*; este dato se ofrece en la tabla *resumen del procesamiento de los casos*, la cual no hemos incluido aquí). Como la variable dependiente tiene tres categorías, tenemos un total de $3 \times 41 = 123$ valores observados con sus correspondientes valores pronosticados (estos 123 valores pueden obtenerse marcando la opción **Probabilidades de casilla** en el subcuadro de diálogo *Regresión logística multinomial: Opciones*). Los estadísticos *Pearson* y *Desvianza*⁶ de la Tabla 6.8 permiten contrastar la hipótesis nula de bondad de ajuste, es decir, la hipótesis de no diferencia entre los valores observados y los pronosticados. Cuanto mayor es el valor de estos estadísticos, peor es el ajuste⁷. Por tanto, los niveles críticos muy pequeños (*sig.* < 0,05) indican que el modelo propuesto no se ajusta bien a los datos.

Los niveles críticos de nuestro ejemplo (*sig.* = 0,505 y *sig.* = 0,537) indican que no existe evidencia de que los valores pronosticados difieran de los observados. Por tanto, no existe evidencia de que el ajuste que se consigue con el modelo propuesto difiera significativamente del ajuste perfecto (es decir, del ajuste modelo saturado).

Tabla 6.8. Estadísticos de ajuste global: bondad de ajuste

	Chi-cuadrado	gl	Sig.
Pearson	73,17	74	,505
Desvianza	72,21	74	,537

⁵ El modelo saturado de un análisis concreto depende del número de patrones de variabilidad. El modelo saturado que utiliza, por defecto, el procedimiento **Regresión logística multinomial** es el modelo que corresponde al número de patrones de variabilidad definidos por los factores y covariables incluidos en el análisis. Puesto que los estadísticos de la Tabla 6.8 se obtienen comparando el modelo propuesto con el modelo saturado, el valor de estos estadísticos depende de cuál sea el modelo saturado de referencia, es decir, de cuál sea el número de patrones de variabilidad. Y este número puede modificarse utilizando las opciones del recuadro **Definir subpoblaciones** del subcuadro de diálogo *Regresión logística multinomial: Estadísticos*. Poder elegir el modelo saturado tiene su utilidad. Por ejemplo, definiendo el mismo modelo saturado (es decir, el mismo número de patrones de variabilidad o *subpoblaciones*) para dos modelos anidados, la diferencia entre las desvianzas de ambos modelos puede utilizarse para valorar la significación simultánea de varios efectos (desde el punto de vista del alejamiento del ajuste perfecto).

⁶ El estadístico de *Pearson* es el mismo que se suele utilizar para contrastar la hipótesis de bondad de ajuste y la hipótesis de independencia en tablas de contingencias: $X^2 = \sum_i (n_i - \hat{m}_i)^2 / \hat{m}_i$ (donde n_i se refiere a los valores observados y $\hat{m}_i$ a los pronosticados). El estadístico *desvianza* es la razón de verosimilitudes que se obtiene al comparar la desvianza del modelo propuesto y la del modelo saturado: $G^2 = 2 [\sum_i n_i \log_e (n_i / \hat{m}_i)]$.

⁷ Con muestras grandes, la distribución de estos estadísticos se aproxima a la distribución ji-cuadrado con un número de grados de libertad que depende del número de ecuaciones estimadas, del número de patrones de variabilidad y del número de coeficientes de regresión estimados; en concreto, $gl = (K - 1)(H - p - 1)$, donde K es el número de categorías de la variable dependiente, H es el número de patrones de variabilidad distintos y p es el número de variables independientes. Para que la aproximación sea aceptable es necesario que haya varios casos por cada patrón de variabilidad; por tanto, no es aconsejable utilizar estos estadísticos si el modelo incluye variables independientes cuantitativas.

En nuestro ejemplo tenemos 2 ecuaciones de regresión, 41 patrones de variabilidad y 3 variables independientes; por tanto, $gl = (3 - 1)(41 - 3 - 1) = 74$.

Significación e interpretación de los coeficientes de regresión

La Tabla 6.9 contiene las estimaciones de los coeficientes del modelo y su significación. Sustituyendo en [6.4] los coeficientes por sus estimaciones obtenemos

$$\begin{aligned}\text{logit}_1 &= 0,91 - 2,20(\text{tto}) + 1,31(\text{sexo}) + 0,38(\text{años}_c) \\ \text{logit}_2 &= 0,89 - 1,86(\text{tto}) - 0,13(\text{sexo}) + 0,11(\text{años}_c)\end{aligned}\quad [6.5]$$

Los coeficientes correspondientes a la variable *tto* son significativamente distintos de cero tanto en el primer logit (*sig.* = 0,002) como en el segundo (*sig.* = 0,012). Los coeficientes correspondientes a la variable *sexo* no alcanzan la significación estadística en ninguno de los dos logit (*sig.* = 0,075 en el primero y *sig.* = 0,850 en el segundo). El coeficiente correspondiente a la variable *años_c* es significativamente distinto de cero en el primer logit (*sig.* < 0,0005) pero no en el segundo (*sig.* = 0,220).

Por tanto, las variables *tto* y *años_c* ayudan a distinguir los pacientes que recaen el *primer año* de los que *no recaen*; y la variable *tto* ayuda a distinguir los pacientes que recaen el *segundo año* de los que *no recaen*. Y dado que los coeficientes asociados a la variable *sexo* no alcanzan la significación estadística, lo apropiado sería eliminar esa variable del análisis para mejorar las estimaciones.

Tabla 6.9. Estimaciones de los parámetros (variables independientes *tto*, *sexo* y *años_c*)

Recaída ^a		B	Error típ.	Wald	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B)	
								L. inferior	L. superior
Primer año	Intersección	,91	,75	1,49	1	,222			
	tto	-2,20	,71	9,64	1	,002	,11	,03	,44
	sexo	1,31	,73	3,17	1	,075	3,70	,88	15,62
	años_c	,38	,10	14,64	1	,000	1,46	1,20	1,77
Segundo año	Intersección	,89	,75	1,41	1	,235			
	tto	-1,86	,74	6,25	1	,012	,16	,04	,67
	sexo	-,13	,71	,04	1	,850	,87	,22	3,52
	años_c	,11	,09	1,51	1	,220	1,12	,94	1,33

^a. La categoría de referencia es: No recae.

- *Coeficientes $\hat{\beta}_{01}$ y $\hat{\beta}_{02}$.* La intersección es, en ambas ecuaciones, el logit pronosticado cuando las tres variables independientes valen cero, es decir, el logit pronosticado para *tto* = “estándar”, *sexo* = “mujer” y *años_c* = 14 (pues la variable *años_c* está centrada en 14). El signo positivo de ambas intersecciones indica que, en las mujeres con 14 años de consumo que reciben el tratamiento estándar, *recaer durante el primer o el segundo año* es más probable que *no recaer*. Pero las diferencias observadas no alcanzan la significación estadística (*sig.* = 0,222 en el primer logit y *sig.* = 0,235 en el segundo).
- *Coeficientes $\hat{\beta}_{11}$ y $\hat{\beta}_{12}$ (*tto*).* El signo negativo del coeficiente asociado a la variable *tto* en el primer logit (-2,20) indica que la *odds* de recaer el *primer año* aumenta al

disminuir el tratamiento. Aquí hay que tener presentes dos cosas: (1) que *disminuir* el tratamiento significa pasar de 1 a 0, es decir de combinado a estándar, y (2) que las *odds* se están calculando respecto de la categoría *no recaer*, que es la categoría de referencia. El valor exponencial del coeficiente ($e^{-2,20} = 0,11$) permite concretar que la *odds* de recaer el *primer año* con el tratamiento combinado es un 11% de la *odds* de recaer con el tratamiento estándar⁸.

En el segundo logit ocurre algo parecido. El signo negativo del coeficiente ($-1,86$) indica que la *odds* de recaer el *segundo año* aumenta con el tratamiento estándar. Y el valor exponencial del coeficiente ($e^{-1,86} = 0,16$) permite concretar que la *odds* de recaer el *segundo año* con el tratamiento combinado es un 16% de la *odds* de recaer con el tratamiento estándar.

- Coeficientes $\hat{\beta}_{21}$ y $\hat{\beta}_{22}$ (*sexo*). Puesto que los coeficientes de regresión asociados a la variable *sexo* no alcanzan la significación estadística, no existe evidencia de que esta variable ayude a distinguir los sujetos que recaen de los que no recaen.
- Coeficientes $\hat{\beta}_{31}$ y $\hat{\beta}_{32}$ (*años_c*). El signo positivo del coeficiente correspondiente a la variable *años_c* en el primer logit (0,38) indica que la *odds* de recaer el *primer año* aumenta cuando aumentan los años de consumo. El valor exponencial del coeficiente ($e^{0,38} = 1,46$) permite concretar que la *odds* de recaer el *primer año* va aumentando un 46% con cada año más de consumo. En el segundo logit se observa la misma tendencia, pero el correspondiente coeficiente de regresión no alcanza a ser significativamente distinto de cero.

El estadístico de Wald (Tabla 6.9) sirve para valorar la significación estadística de *cada coeficiente de regresión* (en el ejemplo, dos coeficientes por variable independiente). La razón de verosimilitudes (Tabla 6.10, columna *chi-cuadrado*) sirve para valorar la significación estadística asociada a *cada variable independiente*. Cuando el valor de un coeficiente es grande también tiende a serlo su error típico; en estos casos, el estadístico de Wald se vuelve conservador y es preferible valorar la significación estadística con la razón de verosimilitudes (Hauck y Donner, 1977; Jennings, 1986).

La primera columna de la Tabla 6.10 ofrece el valor que toma la desviación al eliminar cada variable del modelo propuesto (-2 *verosimilitud del modelo reducido*). Por ejemplo, la desviación del modelo que incluye todos los efectos menos *tto* vale 107,72. La razón de verosimilitudes se calcula comparando esta desviación con la del modelo que incluye todos los efectos, la cual sabemos que vale 95,21 (ver Tabla 6.6, *modelo final*). Por ejemplo, siendo $-2LL_{\text{final}}$ la desviación del modelo propuesto y $-2LL_{\text{reducido}}$ la desviación de ese mismo modelo sin la variable *tto*, la razón de verosimilitudes asociada al efecto de la variable *tto* vale

$$G_{\text{tto}}^2 = -2LL_{\text{reducido}} - (-2LL_{\text{final}}) = 107,72 - 95,21 = 12,51$$

⁸ Si los tratamientos se codificaran al revés (*combinado* = 0 y *estándar* = 1) el coeficiente estimado para la variable *tto* es positivo (2,20) y su valor exponencial vale aproximadamente 9. Esto significa que la *odds* de recaer el *primer año* con el tratamiento *estándar* es nueve veces la *odds* de recaer con el tratamiento *combinado* (a este resultado puede llegarse simplemente calculando el inverso del valor exponencial del coeficiente: $1/0,11 = 9,10$).

El nivel crítico asociado a 12,51 ($\text{sig.} = 0,002$) permite rechazar la hipótesis de que el efecto de la variable *tto* es nulo. Y el rechazo de esta hipótesis permite concluir que la variable *tto* contribuye a reducir el desajuste del modelo nulo. Lo mismo vale decir de la variable *años_c* ($\text{sig.} < 0,0005$), pero no de la variable *sexo*, cuyo efecto no alcanza la significación estadística (pues $\text{sig.} = 0,072$ es mayor que 0,05).

Tabla 6.10. Desvianzas y razones de verosimilitudes al eliminar cada efecto

Efecto	Criterio de ajuste del modelo	Contrastes de la razón de verosimilitud		
	-2 log verosimilitud del modelo reducido	Chi-cuadrado	gl	Sig.
Intersección	97,19	1,98	2	,371
tto	107,72	12,51	2	,002
sexo	100,46	5,26	2	,072
años_c	119,20	23,99	2	,000

Pronósticos y clasificación

Al igual que en la regresión logística binaria, los pronósticos de la regresión nominal pueden utilizarse para clasificar los casos (también para obtener los residuos y, con ello, según veremos, una nueva forma de evaluar la calidad del modelo). La clasificación se realiza a partir de las probabilidades pronosticadas. Y éstas se obtienen mediante

$$\hat{\pi}_k = \frac{e^{\text{logit}_k}}{\sum_k e^{\text{logit}_k}} \quad [6.6]$$

Por ejemplo, a un hombre ($\text{sexo} = 1$) con 14 años de consumo ($\text{años_c} = 0$) que ha recibido el tratamiento combinado ($\text{tto} = 1$) le corresponden, aplicando [6.5], los siguientes pronósticos:

$$\begin{aligned} \text{logit}_1 &= 0,91 - 2,20(1) + 1,31(1) + 0,38(0) = 0,02 & \rightarrow & e^{0,02} = 1,02 \\ \text{logit}_2 &= 0,89 - 1,86(1) - 0,13(1) + 0,11(0) = -1,10 & \rightarrow & e^{-1,10} = 0,33 \\ \text{logit}_3 &= 0 & \rightarrow & e^0 = 1 \end{aligned}$$

Aplicando ahora [6.6]:

$$\begin{aligned} \hat{\pi}_1 &= 1,02 / (1,02 + 0,33 + 1) = 0,43 \\ \hat{\pi}_2 &= 0,33 / (1,02 + 0,33 + 1) = 0,14 \\ \hat{\pi}_3 &= 1 / (1,02 + 0,33 + 1) = 0,42 \end{aligned}$$

Estos pronósticos son los que ofrece el SPSS al marcar la opción **Probabilidades de respuesta estimadas** del subcuadro de diálogo *Guardar*. La clasificación que recoge la Tabla 6.11 se basa en estos pronósticos. Las filas de la tabla clasifican los casos por su

valor *observado* (su valor en la variable dependiente); las columnas, por su valor *pronosticado* (la categoría con la probabilidad estimada más alta).

En la diagonal principal se encuentran los casos bien clasificados (60); fuera de la diagonal, los mal clasificados (24). La última columna informa del porcentaje de casos correctamente clasificados en cada categoría y en total. El porcentaje de clasificación correcta alcanza el 71,4%, aunque no todas las categorías se pronostican igual de bien: el porcentaje de clasificación correcta oscila entre el 20% de la segunda categoría y el 90,5% de la primera.

Al interpretar el porcentaje de casos correctamente clasificados debe tenerse en cuenta que un buen modelo desde el punto de vista de los pronósticos que ofrece puede no ser un buen modelo desde el punto de vista de su capacidad para clasificar casos correctamente (y al revés). Una tabla de clasificación no contiene información acerca de cómo se distribuyen las probabilidades asignadas a cada grupo, es decir, no contiene información acerca de si las probabilidades individuales en las que se basa la clasificación son muy distintas o se parecen. Y, obviamente, no es lo mismo clasificar a un sujeto cuando las probabilidades pronosticadas para cada categoría valen, por ejemplo, 0,95, 0,20 y 0,10, que clasificarlo cuando esas probabilidades valen, por ejemplo, 0,43, 0,14 y 0,42 (como las probabilidades calculadas más arriba). En ambos casos el sujeto sería asignado a la primera categoría, pero en el primer caso se tendría mayor confianza en que la clasificación está bien hecha.

También conviene recordar que el porcentaje de casos correctamente clasificados únicamente debe utilizarse como un criterio de ajuste cuando el objetivo del análisis sea clasificar casos. Si el objetivo del análisis es identificar las variables que contribuyen a entender o explicar el comportamiento de la variable dependiente, es preferible utilizar medidas de ajuste tipo R^2 (ver Hosmer y Lemeshow, 2000, págs. 156-160).

Tabla 6.11. Resultados de la clasificación

Observado	Pronosticado			
	Primer año	Segundo año	No recae	% correcto
Primer año	38	1	3	90,5%
Segundo año	6	3	6	20,0%
No recae	8	0	19	70,4%
Porcentaje global	61,9%	4,8%	33,3%	71,4%

Interacción entre variables independientes

En el capítulo anterior sobre regresión logística binaria hemos explicado cómo interpretar los coeficientes asociados a la interacción entre variables independientes (*covariables* en el contexto de la regresión logística binaria). Lo dicho allí sirve aquí. Lo único que cambia es la forma que cada procedimiento tiene de incluir los términos relativos al efecto de la interacción. En el procedimiento **Regresión logística multinomial** hay que recurrir a las opciones que ofrece el subcuadro de diálogo *Modelo*.

Regresión por pasos

Al igual que en otros tipos de regresión, también en la regresión logística nominal es posible proceder por pasos para construir un modelo de regresión. Cuando no se tiene una hipótesis concreta acerca de las relaciones subyacentes entre las variables estudiadas, proceder por pasos puede ayudar a encontrar el modelo que mejor describe esas relaciones. El objetivo de la regresión por pasos es encontrar el modelo capaz de ofrecer el mejor ajuste con el menor número de términos.

Todo lo dicho en el capítulo anterior sobre la regresión por pasos sirve también aquí (en caso necesario, revisar el apartado *Regresión jerárquica o por pasos* del capítulo anterior). Pero el ajuste por pasos del procedimiento **Regresión logística multinomial** posee algunas peculiaridades que conviene señalar.

En primer lugar, el ajuste por pasos no se solicita en el cuadro de diálogo principal, como en el procedimiento **Regresión logística binaria**, sino en el subcuadro de diálogo *Regresión logística multinomial: Modelo*. Las listas de selección de variables de este subcuadro de diálogo permiten elegir entre (1) construir un modelo en *un único paso* (llevando las variables a la lista **Términos de entrada forzada**) y (2) construir un modelo *por pasos* (llevando las variables a la lista **Términos de pasos sucesivos**). En ambos casos es necesario haber elegido previamente la opción **Personalizado/Pasos sucesivos**.

En segundo lugar, la forma de incluir o eliminar términos difiere de las que hemos estudiado hasta ahora. En regresión logística binaria (también en regresión lineal), los métodos por pasos funcionan incorporando o eliminando variables una a una o por bloques. En regresión logística multinomial no es posible definir bloques de variables. Y la incorporación y eliminación de variables se hace respetando el **principio de jerarquía**: si un modelo incluye un término de orden superior, también debe incluir todos los términos de orden inferior que forman parte de él (por ejemplo, si un modelo incluye la interacción X_1X_2 , también debe incluir los efectos principales de X_1 y X_2).

Sobredispersión

El problema de la sobredispersión ya lo hemos tratado en el capítulo anterior, en el apartado *Dispersión proporcional a la media*. Para todo lo relativo al concepto de sobredispersión y a las consecuencias que se derivan de ella, lo dicho allí sirve también aquí. El concepto de sobredispersión sigue siendo el mismo y sus consecuencias también. Y también aquí se sigue utilizando un *parámetro de escala* para cuantificar el grado de dispersión.

El parámetro de escala puede estimarse dividiendo la desviación del modelo propuesto entre sus grados de libertad. Cuando la dispersión observada y la esperada son iguales, ese cociente toma un valor en torno a 1; un resultado mayor que 1 indica sobredispersión (valores mayores que 2 son problemáticos); un resultado menor que 1 indica infradispersión (la infradispersión es infrecuente).

La desviación y los grados de libertad necesarios para estimar el parámetro de escala son los que el procedimiento **Regresión logística multinomial** ofrece en la tabla de

estadísticos de bondad de ajuste (ver Tabla 6.8). En nuestro ejemplo, el cociente entre la desviación (72,21) y sus grados de libertad (74) vale 0,98, es decir un valor próximo a 1 que indica que, con el modelo propuesto, no parece que el grado de dispersión sea un problema.

Ya hemos señalado que los efectos indeseables de la sobredispersión pueden atenuarse aplicando una sencilla corrección a los errores típicos de los coeficientes. La corrección consiste en multiplicar cada error típico por la raíz cuadrada del valor estimado para el parámetro de escala (en nuestro ejemplo, por la raíz cuadrada de 0,98).

El procedimiento **Regresión logística multinomial** ofrece la posibilidad de corregir automáticamente la dispersión observada aplicando bien una estimación del parámetro de escala basada en los datos (0,98 en nuestro ejemplo), o bien un valor concreto fijado por el usuario. Estas opciones están disponibles en el menú desplegable **Escala** del subcuadro de diálogo *Regresión logística multinomial: Opciones*.

Regresión ordinal

Las variables categóricas *ordinales* son variables cuyas categorías poseen un orden natural (están cuantitativamente ordenadas). En el ámbito de las ciencias sociales y de la salud es frecuente encontrarse con este tipo de variables. Por ejemplo: la gravedad de un síntoma o de una enfermedad (leve, moderada, severa); el grado de satisfacción con un tratamiento (muy insatisfecho, insatisfecho, satisfecho, muy satisfecho); la opinión o actitud que se tiene sobre una determinada cuestión (muy desfavorable, desfavorable, indiferente, favorable, muy favorable); etc.

Una respuesta categórica ordinal podría analizarse aplicando un modelo de regresión logística nominal (ver ecuación [6.1]), pero con un modelo de estas características se estaría pasando por alto el hecho de que las categorías de la variable se encuentran cuantitativamente ordenadas.

El modelo de regresión ordinal

La regresión logística binaria puede adaptarse para incorporar las propiedades de una variable dependiente ordinal. Esta adaptación puede hacerse aplicando diferentes estrategias (ver, por ejemplo, Hosmer y Lemeshow, 2000, págs. 288-291; o McCullag, 1980), pero la más habitual consiste en modelar, no la *probabilidad individual* de cada categoría, sino la *probabilidad acumulada* hasta cada categoría.

Seguimos trabajando con un modelo lineal generalizado con función de enlace logit y, al igual que en el caso de la regresión logística nominal, seguimos utilizando $K - 1$ ecuaciones (K se refiere al número de categorías de la variable dependiente). Pero ahora, en cada ecuación no se está comparando la probabilidad de pertenecer a una categoría concreta con la de pertenecer a la categoría de referencia, sino la probabilidad de pertenecer a una categoría o a las que tienen un código menor que ella con la probabilidad de pertenecer a las categorías con códigos mayores que ella.

En nuestro ejemplo sobre 84 pacientes con problemas de adicción al alcohol (archivo *Tratamiento adicción alcohol*) hay una variable politómica llamada *recaída* que ya hemos utilizado en este mismo capítulo al estudiar la regresión logística nominal. Recordemos que la variable *recaída* tiene tres categorías: el código 1 corresponde a los pacientes que recaen durante el primer año tras finalizar el tratamiento; el código 2, a los que recaen durante el segundo año; el código 3, a los que no recaen en los dos primeros años. Un código menor indica una recaída más temprana. La Tabla 6.1 ofrece un resumen de esta variable combinada con la variable *tto* (tratamiento).

Aunque la variable *recaída* ya la hemos analizado aplicando un modelo de regresión *nominal*, las características de la variable (categorías ordenadas según el tiempo que se tarda en recaer) permiten utilizarla como variable dependiente de un modelo de regresión *ordinal*. Para ello, puesto que la variable tiene $K = 3$ categorías, es necesario definir $K - 1 = 2$ ecuaciones. Cada una de estas ecuaciones modela una *odds* particular:

$$\begin{aligned} odds_1 &= P(recaída = 1) / P(recaída > 1) \\ odds_2 &= P(recaída \leq 2) / P(recaída > 2) \end{aligned} \quad [6.7]$$

En *odds*₁ se está comparando la probabilidad de pertenecer a la primera categoría (código 1) con la de pertenecer a las otras dos categorías (códigos mayores que 1, es decir, códigos 2 y 3). En *odds*₂ se está comparando la probabilidad de pertenecer a las dos primeras categorías (códigos 1 y 2) con la de pertenecer a la tercera categoría (códigos mayores que 2, es decir, código 3). En ambas *odds* intervienen todas las categorías de la variable dependiente.

Con una variable dependiente con K categorías y siendo $F_k = P(Y \leq k)$, es posible definir $K - 1$ *odds acumuladas* no redundantes del tipo

$$odds_k = F_k / (1 - F_k) \quad [6.8]$$

Y el modelo de regresión logística ordinal que permite modelar estas *odds* adopta la siguiente forma:

$$\log_e(odds_k) = \beta_{0k} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_j X_j + \cdots + \beta_p X_p \quad [6.9]$$

(con $k = 1, 2, \dots, K - 1$; p se refiere al número de variables independientes). A pesar de que en [6.9] se están definiendo $K - 1$ ecuaciones de regresión, el único término de la parte derecha que cambia de una ecuación a otra es β_{0k} (es el único término que aparece acompañado del subíndice k). A este término se le suele llamar *umbral* (*threshold*) y, por lo general, carece de interés. Cada ecuación tiene su propio umbral; es un valor que interviene en los pronósticos del modelo (igual que la constante o intersección en una ecuación de regresión lineal) pero que no contiene información sobre la relación entre las variables. Puesto que las probabilidades acumuladas (es decir, los numeradores de [6.8]) son crecientes en k , β_{0k} también lo es.

Los coeficientes β_1 a β_p son los mismos en las $K - 1$ ecuaciones definidas en [6.9] (es decir, no cambian de una ecuación a otra). Por tanto, en un modelo de regresión logística basado en las probabilidades acumuladas se está asumiendo que el efecto de ca-

da variable independiente sobre la dependiente es constante (es el mismo) en todas las ecuaciones. De ahí que al modelo propuesto en [6.9] se le llame modelo de *odds proporcionales* (una característica útil e interesante pero que, según veremos, será necesario chequear).

Una variable independiente (regresión simple)

Veamos cómo utilizar el SPSS para modelar las *odds* propuestas en [6.7] mediante una ecuación logística ordinal. Vamos a comenzar con el caso más simple, es decir, con una sola variable independiente. Para ajustar un modelo de regresión logística ordinal con *recaída* como variable dependiente y *tto* como variable independiente (seguimos utilizando el archivo *Tratamiento adicción alcohol*, el cual puede descargarse de la página web del manual):

- Seleccionar la opción **Regresión > Ordinal** del menú **Analizar** para acceder al cuadro de diálogo *Regresión ordinal* y trasladar la variable *recaída* al cuadro **Dependiente** y la variable *tto* a la lista **Factores**⁹.

Aceptando estas selecciones se obtienen, entre otros, los resultados que muestran las Tablas 6.12 a 6.15.

Ajuste global

Las Tablas 6.12 a 6.14 contienen la información necesaria para valorar el ajuste global del modelo. La información de la Tabla 6.12 permite contrastar la hipótesis nula de que el conjunto de variables independientes incluidas en el análisis (de momento, solo *tto*) no contribuyen a reducir el desajuste del modelo nulo. La tabla incluye la desviación del modelo nulo (*sólo la intersección*: $-2LL_0 = 31,55$), la desviación del modelo propuesto (*final*: $-2LL_1 = 17,19$) y la diferencia entre ambas, es decir, la razón de verosimilitudes G^2 (*chi-cuadrado*):

$$G^2 = -2LL_0 - (-2LL_1) = 31,55 - 17,19 = 14,36$$

Este estadístico sirve para contrastar la hipótesis nula de que los términos en que difieren el modelo nulo y el modelo propuesto valen cero en la población. El rechazo de esta hipótesis indica que el modelo propuesto contribuye a reducir el desajuste del modelo nulo. En nuestro ejemplo, el nivel crítico asociado a G^2 (*sig.* < 0,0005) permite rechazar la hipótesis de que el coeficiente de regresión correspondiente a la variable *tto* vale cero en la población y, consecuentemente, se puede concluir que la variable *tto* contribuye a reducir el desajuste del modelo nulo.

⁹ Puesto que la variable *tto* es dicotómica, puede definirse indistintamente como *factor* y como *covariable*. De ambas formas se obtiene el mismo resultado, pero hay que vigilar, en la interpretación, cuál es la categoría de referencia (pues la *odds ratio* puede calcularse tanto dividiendo *estándar* entre *combinado* como *combinado* entre *estándar*).

Los estadísticos de la Tabla 6.13 permiten valorar el ajuste global del modelo a partir de la comparación entre los valores observados y los pronosticados (en caso necesario, revisar la explicación ofrecida a propósito de la Tabla 6.8). Los niveles críticos del ejemplo ($\text{sig.} = 0,112$ y $\text{sig.} = 0,110$) indican que no existe evidencia de que los valores pronosticados difieran de los observados. Estos estadísticos solo deben utilizarse con muestras grandes.

Por último, los estadísticos tipo R^2 que ofrece la Tabla 6.14 permiten cuantificar en qué medida el modelo propuesto consigue reducir el desajuste del modelo nulo. El estadístico de Nagelkerke indica que la variable *tto* consigue reducir ese desajuste en un 18%.

Tabla 6.12. Estadísticos de ajuste global: desviación y razón de verosimilitudes

Modelo	-2 log de la verosimilitud	Chi-cuadrado	gl	Sig.
Sólo intersección	31,55			
Final	17,19	14,36	1	,000

Tabla 6.13. Estadísticos de ajuste global: bondad de ajuste

	Chi-cuadrado	gl	Sig.
Pearson	2,53	1	,112
Desviación	2,56	1	,110

Tabla 6.14. Estadísticos de ajuste global: pseudo R -cuadrado

Cox y Snell	,16
Nagelkerke	,18
McFadden	,08

Significación e interpretación de los coeficientes de regresión

Para terminar, la Tabla 6.15 ofrece las estimaciones de los coeficientes de regresión. Puesto que la variable dependiente, *recaída*, tiene tres categorías, la tabla ofrece dos ecuaciones de regresión (las dos ecuaciones definidas en [6.7]).

La etiqueta *umbral* se refiere a las intersecciones; la etiqueta *ubicación* se refiere a los coeficientes asociados a las variables independientes (de momento, solo *tto*). Las ecuaciones que ofrece la tabla solo difieren en los *umbrales*; la *ubicación* es la misma en ambas:

$$\begin{aligned}\log_e(\text{odds}_1) &= -0,86 - 1,63 (\text{estándar}) \\ \log_e(\text{odds}_2) &= 0,01 - 1,63 (\text{estándar})\end{aligned}\quad [6.10]$$

Puesto que hemos elegido incluir la variable *tto* como un *factor* (siendo *tto* una variable dicotómica también podríamos haberla incluido como una *covariable*), el algoritmo

de estimación fija en cero el coeficiente de regresión de la categoría *combinado*, que es la categoría con el código más alto ($tto = 1$), y solamente estima el coeficiente de regresión de la categoría *estándar* ($tto = 0$). Tanto el estadístico de Wald como el correspondiente intervalo de confianza indican que el coeficiente asociado al tratamiento estándar ($-1,63$) es significativamente distinto de cero ($sig. < 0,0005$).

El signo negativo del coeficiente de regresión indica que el aumento en la variable *tto* va acompañado de una disminución en el pronóstico lineal (es decir, de una disminución en el logaritmo de la correspondiente *odds*). Para interpretar esto correctamente hay que tener en cuenta dos cosas: (1) un coeficiente negativo indica que las categorías con códigos bajos en la variable dependiente tienden a ser más probables que las categorías con códigos altos; (2) como la categoría que se ha fijado en cero es $tto = 1$ (tratamiento combinado), *aumentar* la variable *tto* significa pasar de combinado a estándar. Por tanto, la recaída (es decir, las categorías con códigos más bajos en la variable dependiente) es más probable con el tratamiento estándar que con el combinado.

Para precisar cuánto más probable es la recaída con el tratamiento estándar hay que obtener el valor exponencial del coeficiente. Pero, para hacer esto, hay que tener presente una característica particular de las *odds* basadas en *probabilidades acumuladas*: cuando las categorías con código menor son más probables que las categorías con código mayor, las *odds* de las primeras respecto de las segundas, toman valores mayores que uno; por el contrario, cuando las categorías con código menor son menos probables que las categorías con código mayor, las *odds* de las primeras respecto de las segundas toman valores menores que uno. Ahora bien, a un pronóstico lineal más alto le corresponde una *odds* más alta; y, sin embargo, mientras que un pronóstico lineal alto apunta hacia las categorías de la variable dependiente que tienen códigos más altos, una *odds* mayor que uno apunta hacia las categorías con códigos más bajos. En consecuencia, para poder interpretar de la forma convencional el valor exponencial de un coeficiente de regresión ordinal (es decir, para poder interpretar que los valores más altos indican que las categorías con códigos más altos son más probables), hay que calcular el valor exponencial cambiando el signo al coeficiente de regresión.

En nuestro ejemplo, el valor exponencial del coeficiente de regresión asociado a la variable *tto* es $e^{1,63} = 5,10$. Este valor indica que la *odds* de las categorías con código menor respecto de las categorías con código mayor aumentan 5,10 veces al cambiar del tratamiento combinado al estándar. Y se asume que este aumento es el mismo en las dos *odds* modeladas.

Tabla 6.15. Estimaciones de los parámetros

		Estimación	Error típ.	Wald	gl	Sig.	Intervalo de confianza 95%	
							L. inferior	L. superior
Umbral	[recaída = 1]	-,86	,32	7,29	1	,007	-1,49	-,24
	[recaída = 2]	,01	,30	,00	1	,976	-,58	,60
Ubicación	[tto=0]	-1,63	,45	13,35	1	,000	-2,51	-,76
	[tto=1]	0 ^a	.	.	0	.	.	.

a. Este parámetro se fija en cero porque es redundante.

Más de una variable independiente (regresión múltiple)

Veamos ahora cómo ajustar e interpretar un modelo de regresión ordinal añadiendo algunas variables independientes. El objetivo del análisis es modelar el logaritmo de las *odds* definidas en [6.7] a partir del *tratamiento* recibido, del *sexo* y del número de *años consumiendo*. Para ajustar un modelo de regresión ordinal con *recaída* como variable dependiente y *tto*, *sexo* y *años_c* como variables independientes:

- Seleccionar la opción **Regresión > Ordinal** del menú **Analizar** para acceder al cuadro de diálogo *Regresión ordinal* y trasladar la variable *recaída* al cuadro **Dependiente** y las variables *tto*, *sexo* y *años_c* a la lista **Covariables**¹⁰.

Aceptando estas selecciones se obtienen, entre otros, los resultados que muestra la Tabla 6.16. La tabla contiene las estimaciones de los coeficientes de las dos ecuaciones de regresión. Recordemos que ambas ecuaciones tienen distinto *umbral* pero la misma *ubicación*:

$$\begin{aligned}\log_e(odds_1) &= -3,58 + 1,56(tto) - 1,07(sexo) - 0,26(años_c) \\ \log_e(odds_2) &= -2,40 + 1,56(tto) - 1,07(sexo) - 0,26(años_c)\end{aligned}\quad [6.11]$$

Tanto el estadístico de Wald como los correspondientes intervalos de confianza indican que los coeficientes asociados a las variables *tto*, *sexo* y *años_c* son, todos ellos, significativamente distintos de cero (*sig.* < 0,05 en los tres casos).

El signo positivo del coeficiente asociado a la variable *tto* (1,56) está indicando que, al aumentar la variable *tto*, aumentan los códigos de la variable *recaída*. Puesto que *aumentar* la variable *tto* significa pasar de 0 a 1 (de estándar a combinado) y *aumentar* los códigos de la variable *recaída* indica menos recaída, el valor positivo del coeficiente asociado a *tto* indica que al pasar del tratamiento estándar al combinado disminuye la recaída. El valor exponencial del coeficiente, $e^{-1,56} = 0,21$ (recordar que para obtener el valor exponencial del coeficiente hay que cambiarle el signo), indica cómo cambian las *odds* de las categorías con código menor respecto de las categorías con código mayor;

Tabla 6.16. Estimaciones de los parámetros

		Estimación	Error típ.	Wald	gl	Sig.	Intervalo de confianza 95%	
							L. inferior	L. superior
Umbral	[recaída = 1]	-3,58	1,03	11,99	1	,001	-5,61	1,55
	[recaída = 2]	-2,40	,99	5,91	1	,015	-4,34	-,46
Ubicación	tto	1,56	,49	9,92	1	,002	,59	2,52
	sexo	-1,07	,52	4,27	1	,039	-2,08	-,06
	años_c	-,26	,06	15,98	1	,000	-,39	-,13

¹⁰ Las variables independientes *categorías* deben ser tratadas como *factores*; las *cuantitativas*, como *covariables*. Las variables *dicotómicas* pueden ser tratadas indistintamente como *factores* y como *covariables*. Ya hemos visto en el apartado anterior cómo se interpreta un variable dicotómica (*tto*) cuando se define como un *factor*; en este apartado vamos a ver cómo se interpreta cuando se define como una *covariable*. Hay detalles que cambian.

en concreto, esas *odds* disminuyen un 79% al pasar del tratamiento estándar al combinado (se asume que esta disminución es la misma en ambas *odds*).

El signo negativo del coeficiente asociado a la variable *sexo* ($-1,07$) está indicando que, al aumentar la variable *sexo*, disminuyen los códigos de la variable *recaída*. Puesto que *aumentar* la variable *sexo* significa pasar de 0 a 1 (de mujer a hombre) y *disminuir* los códigos de la variable *recaída* indica más recaída, el valor negativo del coeficiente asociado a la variable *sexo* indica que la recaída es mayor entre los hombres que entre las mujeres. El valor exponencial del coeficiente, $e^{1,07} = 2,92$, indica cuánto difieren las *odds* de las categorías con código menor de las categorías con código mayor; en concreto, esas *odds* son 2,92 veces mayores entre los hombres que entre las mujeres (se asume que este efecto es idéntico en ambas *odds*).

Por último el signo negativo del coeficiente asociado a la variable *años_c* ($-0,26$) está indicando que, al aumentar los años de consumo, disminuyen los códigos de la variable *recaída* (aumenta la recaída). El valor exponencial del coeficiente, $e^{0,26} = 1,30$, indica que la *odds* de recaer aumenta un 30% con cada año más de consumo (se asume que este efecto es el mismo en ambas *odds*).

Interacción entre variables independientes

En el capítulo anterior sobre regresión logística binaria hemos explicado cómo interpretar los coeficientes asociados a la interacción entre variables independientes. Lo dicho allí sirve también aquí. Lo único que cambia es la forma que cada procedimiento tiene de incluir los términos de interacción. En el procedimiento **Regresión ordinal** hay que recurrir a las opciones que ofrece el subcuadro de diálogo *Modelo*.

Odds proporcionales

El modelo logístico de *probabilidades acumuladas* (ecuación [6.9]) asume que las *odds* definidas en [6.8] son *proporcionales*, es decir, asume que la relación entre las variables independientes y la dependiente es la misma en todas las ecuaciones de regresión. Esto implica que, al estimar los coeficientes de regresión, se está imponiendo la condición de que el resultado debe ser el mismo en todas las ecuaciones. Esto equivale a asumir que las K rectas o planos de regresión (uno por cada categoría de la variable dependiente) son paralelos.

El supuesto de rectas o planos paralelos puede chequearse averiguando si los coeficientes de regresión son iguales al ajustar un modelo de regresión que les permite variar. Los resultados de la Tabla 6.17 permiten realizar esta comprobación (esta tabla se obtiene marcando la opción **Contraste de líneas paralelas** del subcuadro de diálogo *Resultados*). El estadístico $-2LL$ asociado a la *hipótesis nula* (100,57) es la desviación del modelo que asume *odds* proporcionales (rectas o planos paralelos). El estadístico $-2LL$ asociado al modelo *general* (96,67) es la desviación del modelo que no asume *odds* proporcionales. El objetivo del análisis es averiguar si el modelo *general* mejora el ajuste

del modelo que asume *odds* proporcionales. La diferencia entre ambas desviaciones (*chi-cuadrado* = 3,90) permite contrastar la hipótesis nula de que el modelo *general* no reduce el desajuste del modelo que asume *odds* proporcionales. Podrá asumirse que las *odds* son proporcionales cuando la diferencia entre ambas desviaciones sea lo bastante pequeña como para tener asociado un nivel crítico mayor que 0,05.

En nuestro ejemplo, puesto que el nivel crítico (*sig.* = 0,273) es mayor que 0,05, lo razonable es no rechazar la hipótesis nula y, consecuentemente, asumir que las dos *odds* son proporcionales.

Tabla 6.17. Contraste de líneas paralelas

Modelo	-2 log de la verosimilitud	Chi-cuadrado	gl	Sig.
Hipótesis nula	100,57			
General	96,67	3,90	3	,273

Apéndice 6

Funciones de enlace en los modelos de regresión ordinal

Para modelar una variable dependiente ordinal hemos recurrido a una función de enlace *logit*: $\log_e(odds_k)$. Esta función, que es la que el SPSS utiliza por defecto, suele ofrecer buenos resultados con este tipo de variables, particularmente cuando los cambios de una categoría a otra son graduales (no hay categorías especialmente más frecuentes o probables que otras).

Pero la función *logit* no es la única disponible. El procedimiento **PLUM** (opción **Regresión > Ordinal** del menú **Analizar**) ofrece la posibilidad de elegir otras funciones de enlace. Una función que ofrece resultados muy parecidos a la *logit* es la función *probit*: $F^{-1}(\pi_k)$. Esta función reemplaza las probabilidades acumuladas de cada categoría de la variable dependiente por el valor de la curva normal tipificada (puntuación *Z*) que acumula un área igual a esas probabilidades acumuladas. Por tanto, la función de enlace *probit* es útil para modelar variables que se distribuyen normalmente.

La función *log menos log del valor complementario*, es decir, $\log_e(-\log_e(1 - \pi_k))$, es útil para modelar variables en las que la probabilidad acumulada comienza a crecer lentamente desde cero hasta que empieza a aproximarse rápidamente a uno (las categorías con los códigos más altos son más probables que las categorías con los códigos más bajos). Si ocurre lo contrario, es decir, si las categorías con los códigos más bajos son más probables, entonces es preferible utilizar como función de enlace la transformación *log menos log negativa*, $-\log_e(-\log_e(\pi_k))$. Por último, la función *Cauchy inversa*, $\tan[3,1416(\pi_k - 0,5)]$, es apropiada para modelar respuestas con muchos casos extremos.

En las tres funciones del párrafo anterior se está asumiendo que la variable dependiente se distribuye según el modelo de probabilidad multinomial. La diferencia está en si se considera que las categorías más probables son las que tienen los códigos más altos (*log menos log del com-*

plementario), las que tienen los códigos más bajos (*log menos log negativa*) o las categorías con los códigos extremos (*Cauchy inversa*). Todo esto puede apreciarse en las equivalencias que recoge la Tabla 6.18. Todas las funciones de enlace disponibles están expresando la misma idea, pero en distinta escala. Una probabilidad toma valores comprendidos entre cero y uno, y cada valor es simétrico de su complementario (a una probabilidad de 0,25 le corresponde un valor complementario de $1 - 0,25 = 0,75$). Los valores que toman el resto de funciones no tienen ni mínimo ni máximo; y la simetría se pierde en las funciones *log menos log*.

Tabla 6.18. Equivalencia entre distintas funciones de enlace

<i>Prob.</i>	<i>Logit</i>	<i>Probit</i>	<i>L-L negativa</i>	<i>L-L del complem.</i>	<i>Cauchy inversa</i>
0,01	-4,60	-2,33	-1,53	-4,60	-31,82
0,05	-2,94	-1,64	-1,10	-2,97	-6,31
0,10	-2,20	-1,28	-0,83	-2,25	-3,08
0,20	-1,39	-0,84	-0,48	-1,50	-1,38
0,30	-0,85	-0,52	-0,19	-1,03	-0,73
0,40	-0,41	-0,25	0,09	-0,67	-0,32
0,50	0,00	0,00	0,37	-0,37	0,00
0,60	0,41	0,25	0,67	-0,09	0,32
0,70	0,85	0,52	1,03	0,19	0,73
0,80	1,39	0,84	1,50	0,48	1,38
0,90	2,20	1,28	2,25	0,83	3,08
0,95	2,94	1,64	2,97	1,10	6,31
0,99	4,60	2,33	4,60	1,53	31,82

L-L = función log menos log

7

Regresión de Poisson

La **regresión de Poisson** se utiliza para modelar un tipo particular de respuestas llamadas **recuentos**. El término *recuento* tiene aquí un significado especial: *número de ocurrencias de un determinado evento en un periodo de tiempo dado*.

Un recuento es una variable que suele aportar información muy útil en muchos contextos. Ejemplos de este tipo de variables son: el número de episodios depresivos que han experimentado los pacientes de un determinado centro en el último año, el número de cigarrillos/día que fuman los sujetos participantes en un programa de deshabituación, el número de conductas agresivas de un grupo de niños durante un periodo de descanso, el número de accidentes de tráfico que han sufrido los conductores de una ciudad durante los últimos cinco años, etc.

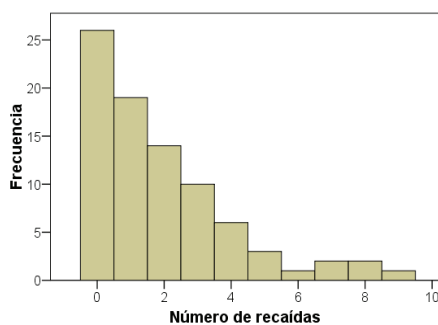
Un recuento no debe confundirse con una **frecuencia**. El término frecuencia lo utilizamos para identificar el número de veces que se repite cada patrón de variabilidad, mientras que un recuento se refiere al valor que toma cada caso individual en una variable cuyos valores reflejan el número de veces que ocurre un evento concreto. Por ejemplo, en una muestra concreta, el número de hombres es una frecuencia; y el número de hombres fumadores con nivel de estudios medios es otra frecuencia. Sin embargo, el número de hijos que tiene cada caso de la muestra es un recuento; y el número de cigarrillos/día que fuma cada caso de la muestra es otro recuento. Por tanto, una frecuencia es un *número de casos*, mientras que un recuento es el *número de veces que un evento de interés se da en un caso*.

En este capítulo veremos cómo modelar recuentos; en el próximo capítulo veremos cómo modelar frecuencias. Ambos tipos de respuestas se modelan aplicando estrategias muy parecidas. Para profundizar en los contenidos de este capítulo recomendamos consultar Cameron y Trivedi (1998), Gardner, Mulvey y Shaw (1995), y Long (1997).

Regresión lineal con recuentos

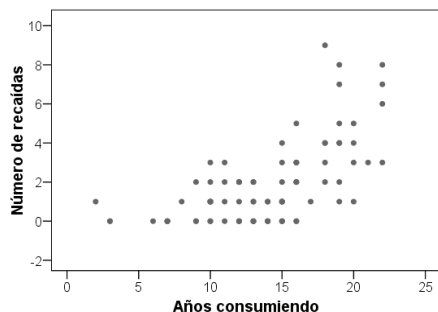
Los recuentos son valores enteros no negativos cuya distribución suele ser asimétrica positiva (los valores más bajos suelen repetirse más que los más altos). La Figura 7.1 muestra una distribución típica de este tipo de datos. Se basa en una muestra de 84 pacientes con problemas de adicción al alcohol. El dato representado es el *número de recaídas* que ha experimentado cada paciente en los dos años siguientes a la finalización de un programa de desintoxicación. El número medio de recaídas es 1,92, con una desviación típica de 2,11.

Figura 7.1. Número de recaídas tras participar en un programa de desintoxicación alcohólica



El *número de recaídas* representado en la en la Figura 7.1 se encuentra en el archivo *Recaídas adicción alcohol*, el cual puede descargarse de la página web del manual. Este archivo también contiene la variable *años* (años consumiendo). La Figura 7.2 muestra el diagrama de dispersión correspondiente a las variables *años consumiendo* y *número de recaídas*. La nube de puntos revela un componente lineal de cierta importancia en la relación entre ambas variables. El coeficiente de correlación de Pearson confirma esta impresión con un valor de 0,63 (*sig.* < 0,0005);

Figura 7.2. Diagrama de dispersión de años consumiendo por número de recaídas



Puesto que el *número de recaídas* es una variable cuantitativa y su relación con *años consumiendo* incluye un claro componente lineal, la relación entre ambas variables podría estudiarse aplicando algún modelo clásico como el análisis de regresión lineal¹. Utilizar esta estrategia implica asumir que el valor esperado de la variable dependiente (μ_i) está linealmente relacionado con la variable independiente (X) :

$$\mu_i = \beta_0 + \beta_1 X_i \quad [7.1]$$

(en caso necesario revisar el Capítulo 2). Los pronósticos que ofrece la ecuación [7.1] forman una línea recta en el plano definido por las variables X e Y . El coeficiente β_0 es la constante o intersección (el punto en el que la recta corta el eje vertical). El coeficiente β_1 refleja la inclinación de la recta respecto del eje horizontal. Cuando no existe relación lineal, la recta es paralela al eje de horizontal ($\beta_1 = 0$).

Al llevar a cabo un análisis de regresión lineal tomando el número de *recaídas* como variable dependiente y los *años consumiendo* como variable independiente se obtienen los resultados que muestra la Tabla 7.1. La recta de regresión resultante es:

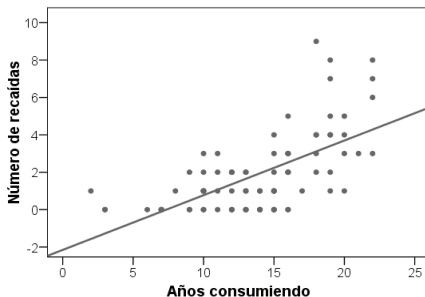
$$\hat{\mu}_i = -2,15 + 0,29(\text{años}) \quad [7.2]$$

La Figura 7.3 muestra esta recta de regresión sobre la nube de puntos de la Figura 7.1. Los resultados de la Tabla 7.1 indican que la relación entre *años consumiendo* y *número de recaídas* es distinta de cero ($\text{sig.} < 0,0005$). Y el valor del coeficiente de regresión

Tabla 7.1. Coeficientes de regresión

Modelo	Coef. no estandarizados		Coef. estandarizados	t	Sig.
	B	Error típ.	Beta		
1 (Constante)	-2,15	,58		-3,69	,000
Años consumiendo	,29	,04	,63	7,33	,000

Figura 7.3. Años consumiendo por número de recaídas con recta de regresión



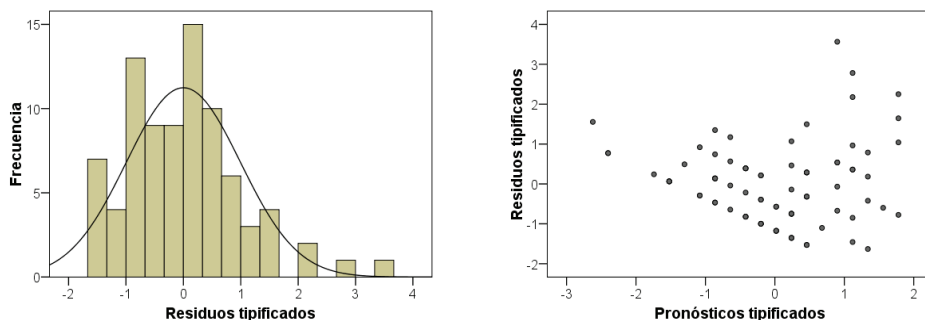
¹ Los recuentos también podrían analizarse aplicando un modelo de regresión logística tras convertirlos en una variable dicotómica (0 = "ocurre", 1 = "no ocurre"). Pero esta estrategia no es del todo apropiada: además de que se estaría perdiendo información, podría ocurrir que el interés del análisis estuviera en pronosticar el número de eventos (regresión de Poisson), no únicamente si el evento ocurre o no (regresión logística).

asociado a la variable *años consumiendo* ($\hat{\beta}_1 = 0,29$) indica que el número estimado de recaídas aumenta 0,29 puntos con cada año más de consumo (una recaída por cada 3,5 años de consumo). El valor del coeficiente de determinación ($R^2 = 0,63^2 = 0,40$) sugiere que el modelo lineal ofrece un buen ajuste a los datos.

Aparentemente, todo ha ido bien con el modelo propuesto en [7.1] y estimado en [7.2]. Sin embargo, sabemos que una ecuación lineal no es la mejor manera de modelar recuentos. Una inspección cuidadosa de la Figura 7.3 permite apreciar varias debilidades en esta estrategia:

1. A pesar de que la variable dependiente no puede tomar valores negativos, la ecuación [7.2] ofrece pronósticos negativos para un amplio rango de valores de la variable independiente (de hecho, únicamente se obtienen pronósticos positivos a partir de 8 años de consumo).
2. La distribución de los residuos no es normal (ver Figura 7.4, gráfico de la izquierda): conserva parte de la asimetría positiva de la variable dependiente.
3. Las varianzas no son homogéneas: la dispersión de la nube de puntos en torno a la recta no es homogénea; va aumentando conforme aumentan los valores de la variable independiente (ver Figura 7.4, gráfico de la derecha).

Figura 7.4. Distribución de los residuos del modelo de regresión lineal



Regresión de Poisson con recuentos

Los resultados del apartado anterior indican que, al analizar recuentos mediante un modelo de regresión lineal, surgen algunos de los problemas que también vimos que surgían con este tipo de modelos al analizar respuestas dicotómicas (ver, en el Capítulo 5, el apartado *La función lineal*). Estos problemas pueden resumirse de la siguiente manera: se corre el riesgo de obtener pronósticos fuera de rango (pronósticos negativos a pesar de que la variable dependiente únicamente puede tomar valores no negativos) y no es fácil conseguir que los errores se distribuyan como se asume que se distribuyen

en un modelo lineal clásico (normalmente y con varianzas homogéneas en cada patrón de variabilidad). Estos problemas derivados de analizar recuentos mediante modelos de regresión lineal obligan a recurrir a otro tipo de estrategias.

El modelo de regresión de Poisson

La más simple y extendida de las estrategias disponibles para analizar recuentos se conoce como **regresión de Poisson**:

$$\mu_i = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p} \quad [7.3]$$

El nombre asignado a esta forma de regresión le viene de la distribución elegida para representar la variabilidad del componente aleatorio (la distribución de Poisson se describe en el Apéndice 1, en el apartado *Distribuciones de la familia exponencial*).

¿Cómo consigue un modelo de estas características superar los problemas asociados a la modelización de recuentos mediante un modelo lineal clásico? En primer lugar, al colocar el predictor lineal como potencia del número e (base de los logaritmos naturales) se elimina la posibilidad de que los pronósticos sean negativos: cuando el número e es elevado al pronóstico generado por el predictor lineal siempre se obtiene un valor no negativo para μ_i .

En segundo lugar, para evitar que la distribución de los errores se aleje de la exigida por un modelo de regresión lineal, se asume que el componente aleatorio (la variabilidad en torno al valor esperado asociado a cada patrón de variabilidad) se distribuye según el modelo de probabilidad de Poisson. Esto implica, en primer lugar, que no es necesario asumir normalidad. Y, en segundo lugar, que tampoco es necesario asumir que las varianzas son homogéneas, pues el tamaño de la varianza de una distribución de Poisson depende del tamaño de su media (de hecho, la varianza de la distribución de Poisson es igual a su media, lo cual resulta especialmente útil para modelar variables cuya dispersión aumenta cuando aumenta la media).

Pero estos no son los únicos beneficios que se obtienen al modelar recuentos con un modelo de regresión de Poisson. Un beneficio adicional tiene que ver con el grado de ajuste que se consigue en comparación con el que se consigue con un modelo de regresión lineal: tratándose de recuentos, los pronósticos en escala logarítmica suelen hacer un mejor seguimiento de la variable dependiente que los pronósticos lineales. Y un beneficio más, especialmente importante al realizar inferencias, tiene que ver con los errores típicos de las estimaciones: en el modelo de regresión de Poisson suelen ser sensiblemente más pequeños que en el modelo de regresión lineal equivalente (es decir, al modelar recuentos, las estimaciones de un modelo de regresión de Poisson son más *eficientes* que las de un modelo de regresión lineal).

Tomando el logaritmo de [7.3] se obtiene la formulación convencional del modelo de regresión de Poisson:

$$\log_e(\mu_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad [7.4]$$

Esta formulación permite apreciar que se trata de un modelo lineal con función de enlace logarítmica, es decir, de un modelo de la familia de los modelos lineales generalizados. Se asume que el componente aleatorio se distribuye según el modelo de probabilidad de Poisson (ver Apéndice 1).

Una variable independiente (regresión simple)

Acabamos de estudiar la relación entre las variables *número de recaídas* y *años consumiendo* (ver Tabla 7.1 y Figura 7.3), pero lo hemos hecho sirviéndonos de un modelo de regresión lineal. En este apartado vamos a estudiar esa misma relación ajustando el siguiente modelo de regresión de Poisson:

$$\log_e(\mu_{\text{recaídas}}) = \beta_0 + \beta_1(\text{años}) \quad [7.5]$$

Este modelo propone que el logaritmo del valor esperado del número de recaídas es función lineal del número de años de consumo. Para ajustar un modelo de regresión de Poisson con la variable *recaídas* (número de recaídas) como variable dependiente y la variable *años* (años consumiendo) como variable independiente (ambas variables se encuentran en el archivo *Recaídas adicción alcohol*, el cual puede descargarse de la página web del manual):

- Seleccionar la opción **Modelos lineales generalizados** del menú analizar para acceder al cuadro de diálogo *Modelos lineales generalizados*.
- En la pestaña **Tipo de modelo**, seleccionar la opción **Loglineal de Poisson** del recuadro **Recuentos** (se obtiene idéntico resultado si en el recuadro **Personalizado** se elige la distribución **Poisson** y la función de enlace **Logaritmo**).
- En la pestaña **Respuesta**, trasladar la variable *recaídas* (número de recaídas) al cuadro **Variable dependiente**.
- En la pestaña **Predictores**, trasladar la variable *años* (años consumiendo) a la lista **Covariables**.
- En la pestaña **Modelo**, trasladar la variable *años* a la lista **Modelo**.
- En la pestaña **Estadísticos**, marcar la opción **Incluir los valores exponenciales de las estimaciones de los parámetros**.

Aceptando estas selecciones se obtienen, entre otros, los resultados que muestran las Tablas 7.2 a 7.4.

Ajuste global: significación estadística

La Tabla 7.2 ofrece varios estadísticos de *ajuste global* (o, mejor, de *desajuste global*, pues eso es lo que miden): todos ellos reflejan el grado en que el ajuste del modelo propuesto se aleja del ajuste de un hipotético modelo que incluyera tantos parámetros como observaciones (el modelo con el mayor ajuste posible). Por tanto, todos ellos toman

un valor tanto mayor cuanto mayor es el desajuste del modelo propuesto (el Apéndice 7 incluye una breve descripción de todos ellos).

Estos estadísticos de *desajuste* no tienen una interpretación directa, pero son muy útiles para comparar modelos rivales. Por ejemplo, la diferencia entre las desviaciones de dos modelos distintos (uno subconjunto del otro) es conceptualmente similar al cambio en R^2 en regresión lineal y se distribuye según ji-cuadrado con los grados de libertad resultantes de restar el número de términos de los dos modelos comparados. Por tanto, la diferencia entre las desviaciones de dos modelos distintos puede utilizarse para valorar la cantidad de ajuste asociada a los términos en que difieren ambos modelos.

En el ejemplo, el efecto de la variable *años* (única variable independiente que incluye el modelo propuesto) puede evaluarse comparando las desviaciones del modelo que incluye esa variable y la del modelo que no la incluye (modelo nulo). La desviación del modelo propuesto vale 107,78 (ver Tabla 7.2); la del modelo nulo, es decir, la del modelo que únicamente incluye la intersección, vale 192,15 (este valor no lo ofrece el SPSS por defecto, pero puede obtenerse ajustando un modelo sin factores ni covariables). La diferencia entre ambas desviaciones vale $192,15 - 107,78 = 84,37$, que es justamente el valor que ofrece la Tabla 7.3 bajo la denominación de *razón de verosimilitudes*.

Esta razón de verosimilitudes sirve para contrastar la hipótesis nula de que todos los términos en que difieren el modelo nulo y el modelo propuesto valen cero en la población. En nuestro ejemplo, el tamaño del nivel crítico (*sig.* $< 0,0005$) permite rechazar esa hipótesis y concluir que el modelo propuesto incluye variables independientes (por ahora, solo *años*) cuyo efecto es distinto de cero. Y esto significa que el modelo propuesto contribuye a reducir el desajuste del modelo nulo.

La Tabla 7.2 contiene un dato adicional de especial interés: el cociente entre la desviación y sus grados de libertad (o entre el estadístico *chi-cuadrado* y sus grados de libertad). Si el modelo propuesto se ajusta bien a los datos, este cociente debe tomar un valor próximo a 1 (volveremos sobre esto más adelante en el apartado *Sobredispersión*).

Tabla 7.2. Estadísticos de bondad de ajuste

	Valor	gl	Valor / gl
Desviación	107,78	82	1,31
Desviación escalada	107,78	82	
Chi-cuadrado de Pearson	98,87	82	1,21
Chi-cuadrado de Pearson escalado	98,87	82	
Log verosimilitud	-132,76		
Criterio de información de Akaike (AIC)	269,53		
AIC corregido para muestras finitas (AICC)	269,68		
Criterio de información bayesiano (BIC)	274,39		
AIC consistente (CAIC)	276,39		

Tabla 7.3. Razón de verosimilitudes

Chi-cuadrado de la razón de verosimilitudes	gl	Sig.
84,37	1	,000

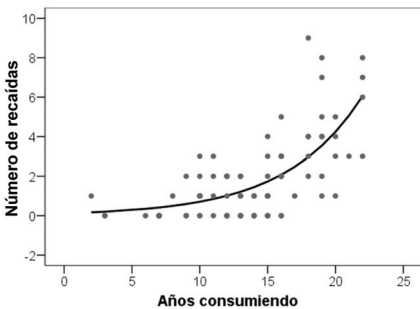
Interpretación de los coeficientes de regresión

El valor de la intersección $(-2,13)$ es el pronóstico que ofrece la ecuación de regresión cuando la variable *años* vale cero. Pero este pronóstico está en escala logarítmica. Para poder interpretarlo hay que devolverlo a su métrica original calculando su valor exponencial: $e^{-2,13} = 0,12$. Este valor aparece en la columna encabezada $\exp(B)$.

Para que este valor tenga algún significado es necesario que el valor cero también tenga algún significado en la variable *años*. Y no es el caso: el valor más pequeño de la variable *años* es 2. Este problema se resuelve utilizando variables *centradas*. Al recodificar la variable *años* asignando el valor cero a los pacientes con 14 años de consumo (el valor de la mediana) se obtiene para la intersección $\hat{\beta}_0$ un valor de 0,37. Este valor es el pronóstico, en escala logarítmica, que ofrece la ecuación de regresión para los pacientes con 14 años de consumo. Y su valor exponencial, $e^{0,37} = 1,45$, es el número de recaídas que la ecuación de regresión pronostica a los pacientes con 14 años de consumo (debe tenerse en cuenta que, aunque la variable dependiente es discreta, los pronósticos de una ecuación de regresión, por lo general, no lo serán).

El coeficiente de regresión de la variable *años* (0,18) refleja cómo cambia el *logaritmo del número estimado de recaídas* por cada unidad que aumenta la variable *años* (un cambio lineal). El valor exponencial del coeficiente ($e^{0,18} = 1,20$) indica cómo cambia el *número estimado de recaídas* por cada unidad que aumenta la variable *años*; en concreto, se estima que el número de recaídas va aumentando un 20% con cada año más de consumo (un cambio no lineal). La Figura 7.5 muestra los pronósticos no lineales (incrementos del 20%) sobre la correspondiente nube de puntos.

Figura 7.5. Años consumiendo por número de recaídas con curva de regresión



En este momento ya tenemos la información necesaria para poder valorar las ventajas de analizar recuentos con un modelo de regresión de Poisson en lugar de hacerlo con un modelo de regresión lineal: (1) los pronósticos no están fuera de rango, (2) no existen problemas con la normalidad de los errores ni con la homogeneidad de sus varianzas, (3) la ecuación hace un mejor seguimiento de la nube de puntos (puede apreciarse comparando las Figuras 7.2 y 7.5) y (4) las estimaciones son más eficientes (los errores típicos de las estimaciones son más pequeños; ver Tablas 7.1 y 7.4).

Una variable independiente dicotómica

Lo dicho en los apartados anteriores a propósito del modelo que incluía una variable independiente cuantitativa (*años*) sirve también para los modelos que incluyen una variable independiente dicotómica. Pero con este tipo de variables hay que prestar atención especial a la interpretación de los coeficientes.

Utilizando la variable *recaídas* (número de recaídas) como variable dependiente y la variable *tto* (tratamiento) como variable independiente (seguimos con el archivo *Recaídas adicción alcohol*, el cual puede descargarse de la página web del manual) se obtienen las estimaciones que muestra la Tabla 7.5:

$$\log_e(\hat{\mu}_{\text{recaídas}}) = \hat{\beta}_0 + \hat{\beta}_1(tto) = 0,94 - 0,71(tto) \quad [7.8]$$

Tanto la intersección como el coeficiente asociado a la variable *tto* son distintos de cero (*sig.* < 0,0005 en ambos casos). Puesto que la variable independiente solamente toma dos valores distintos (0 y 1), la ecuación solo ofrece dos pronósticos distintos:

$$\begin{aligned} \log_e(\hat{\mu}_{\text{recaídas}} | tto = 0) &= 0,94 - 0,71(0) = 0,94 & \rightarrow & e^{0,94} = 2,57 \\ \log_e(\hat{\mu}_{\text{recaídas}} | tto = 1) &= 0,94 - 0,71(1) = 0,23 & \rightarrow & e^{0,23} = 1,26 \end{aligned}$$

La intersección es el pronóstico que se obtiene cuando la variable *tto* vale cero. Por tanto, el valor exponencial del coeficiente (2,57) es el número estimado de recaídas para los pacientes que han recibido el tratamiento *estándar* (*tto* = 0).

El signo negativo del coeficiente de regresión asociado a la variable *tto* indica que la relación entre la variable dependiente y la independiente es negativa: el número estimado de recaídas disminuye cuando aumenta la variable *tto*, es decir, cuando *tto* pasa de 0 a 1 (de estándar a combinado). El valor exponencial del coeficiente (0,49) permite concretar que el número estimado de recaídas con el tratamiento combinado (*tto* = 1) es un 49% del número estimado de recaídas con el estándar (*tto* = 0). De otra forma: el número estimado de recaídas con el tratamiento combinado es un 51% menor que con el estándar.

Puesto que el número estimado de recaídas con el tratamiento estándar es 2,57 (el valor de la intersección), el valor exponencial del coeficiente asociado a la variable *tto* (0,49) permite conocer el número estimado de recaídas con el tratamiento combinado: $0,49(2,57) = 1,26$. Y estos dos valores no son otra cosa que el número medio de recaídas entre los pacientes que han recibido el tratamiento estándar (2,57) y los que han recibido el tratamiento combinado (1,26).

Tabla 7.5. Estimaciones de los parámetros (variable independiente *tto*)

Parámetro	B	Error típico	Interv. confianza de Wald 95%		Contraste de hipótesis			Exp(B)
			Inferior	Superior	Wald	gl	Sig.	
(Intersección)	,94	,10	,76	1,13	96,34	1	,000	2,57
tto	-,71	,17	-1,04	-,38	18,02	1	,000	,49
(Escala)	1							

Una variable independiente politómica

Las variables politómicas (variables categóricas con más de dos categorías) deben incluirse en el análisis como *factores*, no como *covariables*. Con este tipo de variables hay que prestar especial atención a la interpretación de los coeficientes de regresión. El resto de la información que ofrece el procedimiento no cambia.

En el archivo de datos que venimos utilizando (*Recaídas adicción alcohol*) los pacientes han seguido uno de tres regímenes hospitalarios distintos (variable *régimen*): 1 = “Interno”, 2 = “Externo”, 3 = “Domicilio”. La Tabla 7.6 muestra el número medio de recaídas en cada una de estas tres categorías.

Tabla 7.6. Recaídas con cada régimen hospitalario

Media	
Régimen hospitalario	Número de recaídas
Interno	3,52
Externo	1,07
Domicilio	1,08

Utilizando la variable *recaídas* (número de recaídas) como variable dependiente y la variable *régimen* (régimen hospitalario) como *factor* (esto se controla en la pestaña **Predictores**) se obtienen las estimaciones que muestra la Tabla 7.7.

El procedimiento fija en cero el coeficiente correspondiente a la última categoría del factor (esto se indica en una nota a pie de tabla) y estima los coeficientes del resto de categorías por comparación con el que se ha fijado en cero. Por tanto, de los tres coeficientes asociados a la variable *régimen*, el último de ellos (el correspondiente a la categoría *domicilio*) se ha fijado en cero y únicamente se han estimado los de las categorías *régimen* = 1 (interno) y *régimen* = 2 (externo).

Esta estrategia equivale a crear dos variables dicotómicas (X_1 y X_2) a partir de las tres categorías de la variable *régimen*: los pacientes que puntúan 1 en X_1 y 0 en X_2 pertenecen al régimen *interno*; los pacientes que puntúan 0 en X_1 y 1 en X_2 pertenecen al régimen *externo*; los pacientes que puntúan 0 en ambas variables pertenecen al régimen *domiciliario* (categoría de referencia).

Con las estimaciones que ofrece la Tabla 7.7 se pueden construir las ecuaciones necesarias para obtener los pronósticos asociados a cada centro hospitalario. Puede comprobarse que el valor exponencial de los pronósticos no es otra cosa que el número medio de recuperaciones que corresponde a cada régimen hospitalario (ver Tabla 7.6):

$$\begin{aligned}
 \log_e(\hat{\mu}_{\text{recaídas}} \mid \text{régimen} = 1) &= 0,08 + 1,18 = 1,26 & \rightarrow & e^{1,26} = 3,52 \\
 \log_e(\hat{\mu}_{\text{recaídas}} \mid \text{régimen} = 2) &= 0,08 - 0,01 = 0,07 & \rightarrow & e^{0,07} = 1,07 \\
 \log_e(\hat{\mu}_{\text{recaídas}} \mid \text{régimen} = 3) &= 0,08 + 0 = 0,08 & \rightarrow & e^{0,08} = 1,08
 \end{aligned}$$

Los niveles críticos que ofrece la Tabla 7.7 indican que el pronóstico (número medio de recaídas) para *régimen* = 1, es decir, para el régimen interno, difiere significativa-

mente ($\text{sig.} < 0,0005$) del pronóstico para *régimen* = 3, es decir, para el régimen domiciliario (que es el que se está utilizando como categoría de referencia). Sin embargo, no es posible afirmar que el pronóstico para *régimen* = 2, es decir, para el régimen externo, sea distinto ($\text{sig.} = 0,962$) del pronóstico para el régimen domiciliario.

El valor exponencial de la intersección (1,08) es el pronóstico que ofrece la ecuación de regresión cuando el resto de coeficientes vale cero. Por tanto, es el pronóstico correspondiente a la categoría de referencia (*régimen* = 3), es decir, el pronóstico que ofrece la ecuación para los pacientes que han seguido el régimen domiciliario. En la Tabla 7.6 se puede comprobar que el número medio de recaídas con ese régimen es precisamente 1,08.

El valor exponencial del coeficiente de regresión correspondiente al régimen interno (*régimen* = 1) indica que el número estimado de recaídas para los pacientes que han seguido ese régimen es 3,26 veces mayor que el número estimado de recaídas para los pacientes que han seguido el régimen de referencia (el domiciliario). Efectivamente, multiplicando 3,26 por la media obtenida con el régimen domiciliario (1,08) se obtiene la media obtenida con el régimen interno: $3,26(1,08) = 3,52$ (ver Tabla 7.6).

Finalmente, el valor exponencial del coeficiente de regresión correspondiente al régimen externo (*régimen* = 2) indica que el número estimado de recaídas para los pacientes que han seguido ese régimen es un 99% del número estimado de recaídas para los pacientes que han seguido el régimen de referencia (el domiciliario). Efectivamente, multiplicando 0,99 por la media obtenida con el régimen domiciliario (1,08) se obtiene la media obtenida con el régimen externo: $0,99(1,08) = 1,07$ (ver Tabla 7.6).

Tabla 7.7. Estimaciones de los parámetros (variable independiente *régimen hospitalario*)

Parámetro	B	Error típico	Interv. confianza de Wald 95%		Contraste de hipótesis			Exp(B)
			Inferior	Superior	Wald	gl	Sig.	
(Intersección)	,08	,19	-,30	,45	,16	1	,689	1,08
[régimen=1]	1,18	,22	,76	1,60	29,76	1	,000	3,26
[régimen=2]	-,01	,26	-,52	,50	,00	1	,962	,99
[régimen=3]	,00 ^a	.	.	.	.	.	.	1
(Escala)	1,00							

^a. Establecido en cero ya que este parámetro es redundante.

Para obtener la comparación que falta (régimen interno con régimen externo) basta con cambiar la categoría de referencia. Para ello, en la pestaña **Predictores**, el botón **Opciones** ubicado debajo de la lista **Factores** conduce a un subcuadro de diálogo que permite cambiar el orden de las categorías. La opción **Ascendente**, que es la que se encuentra activa por defecto, fija como categoría de referencia la última (domicilio). Elijiendo la opción **Descendente**, la categoría de referencia pasa a ser la primera (interno).

Al proceder de esta manera se obtiene, para el coeficiente de regresión correspondiente al régimen externo (*régimen* = 2), un valor exponencial de 0,303. Este valor indica que el número estimado de recaídas para los pacientes que han seguido el régimen

externo es aproximadamente un 30% del número estimado de recaídas para los pacientes que han seguido el régimen de referencia (que ahora es el interno). Efectivamente, multiplicando 0,303 por la media obtenida con el régimen interno (3,52) se obtiene la media obtenida con el régimen externo: $0,303(3,52) = 1,07$ (ver Tabla 7.6). La diferencia entre ambos pronósticos es estadísticamente significativa ($\text{sig.} < 0,0005$).

Más de una variable independiente (regresión múltiple)

Veamos cómo estimar e interpretar un modelo de regresión múltiple utilizando tres variables independientes: *años_c* (años consumiendo, centrada), *sexo* y *tto* (tratamiento). Es decir, veamos cómo estimar e interpretar el siguiente modelo de regresión:

$$\log_e(\mu_{\text{recaídas}}) = \beta_0 + \beta_1(\text{años_c}) + \beta_2(\text{sexo}) + \beta_3(\text{tto}) \quad [7.9]$$

Seguimos con el archivo *Recaídas adicción alcohol*, el cual puede descargarse de la página web del manual. La variable *años_c* está centrada en 14 años (recordemos que las variables cuantitativas se centran para facilitar la interpretación de la intersección del modelo). Para estimar el modelo propuesto en [7.9]:

- Seleccionar la opción **Modelos lineales generalizados** del menú analizar para acceder al cuadro de diálogo *Modelos lineales generalizados*.
- En la pestaña **Tipo de modelo**, seleccionar la opción **Loglineal de Poisson** del recuadro **Recuentos** (se obtiene idéntico resultado si en el recuadro **Personalizado** se elige la distribución **Poisson** y la función de enlace **Logaritmo**).
- En la pestaña **Respuesta**, trasladar la variable *recaídas* (número de recaídas) al cuadro **Variable dependiente**.
- En la pestaña **Predictores**, trasladar las variables *años_c* (años consumiendo), *sexo* y *tto* (tratamiento) a la lista **Covariables**.
- En la pestaña **Modelo**, trasladar las variables *años_c*, *sexo* y *tto* a la lista **Modelo**.
- En la pestaña **Estadísticos**, marcar la opción **Incluir los valores exponenciales de las estimaciones de los parámetros**.

Aceptando estas selecciones se obtienen, entre otros, los resultados que muestran las Tablas 7.8 y 7.9.

Ajuste global

La razón de verosimilitudes que ofrece la Tabla 7.8 indica en qué medida el modelo propuesto (el modelo que incluye las variables independientes *años_c*, *sexo* y *tto*) consigue reducir el desajuste del modelo nulo (el modelo que únicamente incluye la intersección). La diferencia entre las desviaciones de ambos modelos (el estadístico *razón de verosimilitudes*) vale 92,27. El nivel crítico asociado a este estadístico ($\text{sig.} < 0,0005$)

indica que el modelo propuesto (las variables *años_c*, *sexo* y *tto* tomadas juntas) consigue reducir significativamente el desajuste del modelo nulo.

Tabla 7.8. Razón de verosimilitudes

Chi-cuadrado de la razón de verosimilitudes	gl	Sig.
92,27	3	,000

Significación de los coeficientes de regresión

El modelo de regresión que estamos ajustando incluye la variable dependiente *recaídas* (número de recaídas) y tres variables independientes: *años_c*, *sexo* y *tto* (ver ecuación [7.9]). La primera columna de la Tabla 7.9 contiene las estimaciones de los correspondientes coeficientes de regresión:

$$\log_e(\hat{\mu}_{\text{recaídas}}) = 0,39 + 0,17(\text{años_c}) + 0,25(\text{sexo}) - 0,44(\text{tto}) \quad [7.10]$$

La significación de cada coeficiente se evalúa con el estadístico de *Wald*, el cual ya sabemos que sirve para contrastar la hipótesis nula de que el correspondiente coeficiente de regresión vale cero en la población. En nuestro ejemplo, las variables *años_c* y *tto* tienen asociados coeficientes significativamente distintos de cero (*sig.* < 0,0005 en el primer caso y *sig.* = 0,011 en el segundo). Sin embargo, el coeficiente de regresión asociado a la variable *sexo* no alcanza la significación estadística (*sig.* = 0,155 > 0,05). Por tanto, únicamente las variables *años_c* y *tto* están contribuyendo a reducir el desajuste del modelo nulo.

Tabla 7.9. Estimaciones de los parámetros (variables independientes *años_c*, *sexo* y *tto*)

Parámetro	B	Error típico	Interv. confianza de Wald 95%		Contraste de hipótesis			Exp(B)
			Inferior	Superior	Wald	gl	Sig.	
(Intersección)	,39	,17	,05	,72	5,18	1,00	,023	1,48
años_c	,17	,02	,13	,21	62,00	1,00	,000	1,18
sexo	,25	,18	-,10	,60	2,02	1,00	,155	1,29
tto	-,44	,17	-,77	-,10	6,50	1,00	,011	,65
(Escala)	1,00							

Las variables independientes que no contribuyen a reducir el desajuste conviene eliminarlas del modelo; esto no solo no altera la calidad del modelo sino que ayuda a que las nuevas estimaciones sean más eficientes. Al eliminar la variable *sexo*, la razón de verosimilitudes apenas cambia (pasa de 92,27 a 90,18; ver Tablas 7.8 y 7.10). Y las nuevas estimaciones (ver Tabla 7.11) permiten construir la siguiente ecuación:

$$\log_e(\hat{\mu}_{\text{recaídas}}) = 0,55 + 0,17(\text{años_c}) - 0,40(\text{tto}) \quad [7.11]$$

Tabla 7.10. Razón de verosimilitudes

Chi-cuadrado de la razón de verosimilitudes	gl	Sig.
90,18	2	,000

Tabla 7.11. Estimaciones de los parámetros (variables independientes *años_c* y *tto*)

Parámetro	B	Error típico	Interv. confianza de Wald 95%		Contraste de hipótesis			Exp(B)
			Inferior	Superior	Wald	gl	Sig.	
(Intersección)	,55	,12	,31	,79	20,45	1,00	,000	1,74
años_c	,17	,02	,13	,21	62,80	1,00	,000	1,18
tto	-,40	,17	-,74	-,07	5,60	1,00	,018	,67
(Escala)	1,00							

Interpretación de los coeficientes de regresión

Al igual que en el resto de modelos de regresión estudiados, el signo de los coeficientes refleja el sentido (positivo o negativo) de la relación entre cada variable independiente y la variable dependiente. Y el valor exponencial de los coeficientes indica cuánto cambia el número de recaídas (la variable dependiente en su métrica original) por cada unidad que aumenta la correspondiente variable independiente; y esto, cualquiera que sea el valor del resto de variables independientes:

- *Coeficiente $\hat{\beta}_0$* . El valor de la intersección (0,55) es el pronóstico que ofrece la ecuación de regresión cuando todas las variables independientes (en el ejemplo, las variables *años_c* y *tto*) valen cero. Este pronóstico está en escala logarítmica. Su valor exponencial ($e^{0,55} = 1,74$) es el número estimado de recaídas para los pacientes que llevan 14 años consumiendo (*años_c* = 0) y que han recibido el tratamiento estándar (*tto* = 0).
- *Coeficiente $\hat{\beta}_1$ (*años_c*)*. El signo positivo del coeficiente de regresión asociado a la variable *años_c* (0,17) indica que el número de recaídas aumenta cuando aumentan los años de consumo. El valor exponencial del coeficiente (1,18) permite concretar que, independientemente del tratamiento recibido, el número estimado de recaídas aumenta un 18% con cada año más de consumo.
- *Coeficiente $\hat{\beta}_2$ (*tto*)*. El signo negativo del coeficiente de regresión asociado a la variable *tto* (-0,40) indica que el número de recaídas disminuye cuando aumenta la variable *tto*, es decir, cuando *tto* pasa de 0 a 1 (de estándar a combinado). El valor exponencial del coeficiente (0,67) permite concretar que, independientemente de los años de consumo, el número estimado de recaídas con el tratamiento combinado es un 67% del estimado con el tratamiento estándar; o bien, que el número estimado de recaídas con el tratamiento combinado es un 33% menor que el estimado con el estándar.

Interacción entre variables independientes

La forma de incorporar a una ecuación de regresión el efecto de la interacción entre variables independientes consiste simplemente en incluir el producto de las variables que interaccionan. Una ecuación de regresión *no aditiva*, con dos variables independientes (X_1 y X_2), adopta la siguiente forma:

$$\log_e(\mu_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \quad [7.12]$$

Para estimar una ecuación de este tipo con el procedimiento **Modelos lineales generalizados** basta con indicar en la pestaña **Modelo** los términos que debe incluir el modelo, a saber, los efectos principales de X_1 y X_2 , y el efecto de la interacción entre X_1 y X_2 .

Al incluir en la ecuación un término con la interacción $X_1 X_2$ la situación se complica bastante. Para facilitar la explicación vamos a considerar tres escenarios: (1) dos variables independientes dicotómicas, (2) dos variables independientes cuantitativas y (3) una variable independiente dicotómica y otra cuantitativa.

Dos variables independientes dicotómicas

Nuestro archivo *Recaídas adicción alcohol* incluye dos variables dicotómicas: *tto* y *sexo*. Una ecuación de regresión no aditiva con el *número de recaídas* como variable dependiente y las variables *tto* y *sexo* como independientes adopta la siguiente forma:

$$\log_e(\mu_{\text{recaídas}}) = \beta_0 + \beta_1(tto) + \beta_2(sexo) + \beta_3(tto \times sexo) \quad [7.13]$$

La Tabla 7.12 muestra los resultados obtenidos al ajustar este modelo. Sustituyendo los parámetros de [7.13] por las estimaciones que ofrece la tabla obtenemos

$$\log_e(\hat{\mu}_{\text{recaídas}}) = 0,99 - 1,57(tto) - 0,07(sexo) + 1,17(tto \times sexo)$$

Únicamente la variable *tto* y la interacción *tto* $\times$ *sexo* tienen asociados coeficientes de regresión significativamente distintos de cero (*sig.* $< 0,05$). No obstante, interpretaremos todos los coeficientes del modelo para aclarar su significado. Para ayudar en la interpretación, la Tabla 7.13 muestra el número medio de recaídas en cada combinación *tto* $\times$ *sexo*.

Tabla 7.12. Estimaciones de los parámetros (variables independientes *tto* y *sexo*)

Parámetro	B	Error típico	Interv. confianza de Wald 95%		Contraste de hipótesis			Exp(B)
			Inferior	Superior	Wald	gl	Sig.	
(Intersección)	,99	,17	,66	1,32	34,33	1	,000	2,69
tto	-1,57	,37	-2,30	-,83	17,55	1	,000	,21
sexo	-,07	,21	-,47	,34	,11	1	,744	,93
tto * sexo	1,17	,42	,35	1,99	7,76	1	,005	3,22
(Escala)	1							

Tabla 7.13. Número medio de recaídas por *tratamiento* y *sexo*

<i>Tratamiento</i>	<i>Sexo</i>	
	Hombres	Mujeres
Estándar	2,52	2,69
Combinado	1,69	0,56

- *Coefficiente $\hat{\beta}_0$* . La intersección es el pronóstico que ofrece la ecuación de regresión cuando todas las variables independientes valen cero. Su valor exponencial (2,69) es el número estimado de recaídas para las mujeres (*sexo* = 0) a las que se les ha administrado el tratamiento estándar (*tto* = 0). En la Tabla 7.13 puede comprobarse que esta estimación no es otra cosa que el número medio de recaídas observado en las mujeres que han recibido el tratamiento estándar.
- *Coefficiente $\hat{\beta}_1$ (*tto*)*. El coeficiente asociado a la variable *tto* recoge el efecto de esa variable cuando *sexo* = 0 (mujeres). El signo negativo del coeficiente (-1,57) indica que el número de recaídas disminuye cuando aumenta la variable *tto*; por tanto, el número estimado de recaídas es menor con el tratamiento combinado (*tto* = 1) que con el estándar (*tto* = 0). El valor exponencial del coeficiente (0,21) permite concretar que, entre las mujeres, el número estimado de recaídas con el tratamiento combinado (0,56; ver Tabla 7.13) es un 21% del número estimado de recaídas con el tratamiento estándar (2,69). Efectivamente, $0,56/2,69 = 0,21$.
- *Coefficiente $\hat{\beta}_2$ (*sexo*)*. El coeficiente asociado a la variable *sexo* recoge el efecto de esa variable cuando *tto* = 0 (estándar). El signo negativo del coeficiente (-0,07) indica que el número de recaídas disminuye al aumentar la variable *sexo*; por tanto, el número de recaídas es menor entre los hombres (*sexo* = 1) que entre las mujeres (*sexo* = 0). El valor exponencial del coeficiente (0,93) permite concretar que, entre quienes han recibido el tratamiento estándar, el número de recaídas entre los hombres (2,52) es un 93% del número de recaídas entre las mujeres (2,69). Efectivamente, $2,52/2,69 = 0,93$. No obstante, esta diferencia no alcanza la significación estadística (*sig.* = 0,774).
- *Coefficiente $\hat{\beta}_3$ (*tto* × *sexo*)*. Por último, el coeficiente de regresión asociado al efecto de la interacción refleja cómo cambia la relación entre el *número de recaídas* y los *tratamientos* en función del *sexo*. Entre los hombres, recaer con el tratamiento combinado respecto de hacerlo con el estándar vale $1,69/2,52 = 0,671$; entre las mujeres, ese cociente vale $0,56/2,69 = 0,208$. El valor exponencial del coeficiente de regresión ($\hat{\beta}_3 = 3,22$) indica que recaer con el tratamiento combinado respecto de hacerlo con el estándar es $0,671/0,208 = 3,22$ veces mayor entre los hombres que entre las mujeres. Exactamente lo mismo vale decir de cómo cambia la relación entre el *número de recaídas* y el *sexo* en función del *tratamiento*².

² $e^{\hat{\beta}_3} = 3,22$ es el factor por el que queda multiplicado $e^{\hat{\beta}_1} = 0,21$ al pasar de *sexo* = 0 (mujeres) a *sexo* = 1 (hombres). También es el factor por el que queda multiplicado $e^{\hat{\beta}_2} = 0,93$ al pasar de *tto* = 0 (estándar) a *tto* = 1 (combinado).

Dos variables independientes cuantitativas

En nuestro archivo *Recaídas adicción alcohol* tenemos las variables cuantitativas *edad* y *años* (años consumiendo). Una ecuación de regresión *no aditiva* con el *número de recaídas* como variable dependiente y la *edad* y los *años consumiendo* como variables independientes adopta la forma:

$$\log_e(\mu_{\text{recaídas}}) = \beta_0 + \beta_1(\text{edad_c}) + \beta_2(\text{años_c}) + \beta_3(\text{edad_c} \times \text{años_c}) \quad [7.14]$$

Recordemos que el coeficiente β_0 únicamente tiene significado cuando también lo tiene el valor cero en todas las variables independientes. Por este motivo, y para facilitar después la interpretación del resto de coeficientes, en lugar de las variables originales *edad* y *años*, la ecuación [7.14] incluye las variables *edad_c* (*edad centrada*) y *años_c* (*años consumiendo centrada*). Ambas variables se han centrado tomando como referencia un valor próximo al centro de sus respectivas distribuciones: 50 en el caso de la *edad* y 14 en el caso de los *años consumiendo*. Por tanto, el valor *edad_c* = 0 se refiere a una *edad* de 50 años y el valor *años_c* = 0 se refiere a 14 años de consumo.

La Tabla 7.14 muestra las estimaciones obtenidas al ajustar el modelo propuesto en [7.14], junto con la significación estadística y los intervalos de confianza asociados a cada coeficiente de regresión:

$$\log_e(\mu_{\text{recaídas}}) = 0,33 - 0,05(\text{edad_c}) + 0,22(\text{años_c}) + 0,00(\text{edad_c} \times \text{años_c})$$

Tabla 7.14. Estimaciones de los parámetros (variables independientes *edad_c* y *años_c*)

Parámetro	B	Error típico	Interv. confianza de Wald 95%		Contraste de hipótesis			Exp(B)
			Inferior	Superior	Wald	gl	Sig.	
(Intersección)	,33	,11	,12	,55	9,15	1	,002	1,40
edad_c	-,05	,02	-,09	-,01	7,39	1	,007	,95
años_c	,22	,03	,17	,28	63,26	1	,000	1,25
edad_c * años_c	,00	,00	,00	,01	1,44	1	,230	1,00
(Escala)	1							

- **Coficiente $\hat{\beta}_0$.** La intersección es el valor que pronostica la ecuación de regresión cuando todas las variables independientes valen cero. Puesto que nuestras variables independientes están centradas en 50 y 14, respectivamente, el valor exponencial de la intersección (1,40) es el número estimado de recaídas para los pacientes que tienen 50 años y que llevan 14 años consumiendo alcohol.
- **Coficiente $\hat{\beta}_1$ (*edad_c*).** El coeficiente asociado a la variable *edad_c* recoge el efecto de esa variable cuando *años_c* = 0 (14 años de consumo). El valor exponencial del coeficiente (0,95) indica que, entre los pacientes con 14 años de consumo, el número estimado de recaídas disminuye un 5% con cada año más de edad.
- **Coficiente $\hat{\beta}_2$ (*años_c*).** El coeficiente asociado a la variable *años_c* recoge el efecto de esa variable cuando *edad_c* = 0 (50 años). El valor exponencial del coe-

ficiente (1,25) indica que, entre los pacientes que tienen 50 años, el número estimado de recaídas va aumentando un 25% con cada año más de consumo.

- **Coefficiente $\hat{\beta}_3$ ($edad_c \times años_c$).** El coeficiente asociado al efecto de la interacción indica cómo cambia la relación entre el número de recaídas y los años de consumo al ir aumentando la edad. Puesto que el nivel crítico obtenido ($sig. = 0,230$) es mayor que 0,05, no puede concluirse que la relación entre el número de recaídas y los años de consumo cambie con la edad.

Una variable independiente dicotómica y una cuantitativa

Consideremos finalmente una ecuación de regresión *no aditiva* con el *número de recaídas* como variable dependiente y las variables *tratamiento* y *años de consumo* como variables independientes:

$$\log_e(\mu_{\text{recaídas}}) = \beta_0 + \beta_1(tto) + \beta_2(años_c) + \beta_3(tto \times años_c) \quad [7.15]$$

En lugar de la variable original *años*, la ecuación [7.15] incluye la variable *años_c*, es decir, la variable *años consumiendo centrada* en 14 (el valor de la mediana). Por tanto, el valor *años_c* = 0 se refiere a 14 años de consumo. La Tabla 7.15 muestra las estimaciones obtenidas al ajustar el modelo propuesto en [7.15]:

$$\log_e(\mu_{\text{recaídas}}) = 0,65 - 0,63(tto) + 0,14(años_c) + 0,08(tto \times años_c)$$

Tabla 7.15. Estimaciones de los parámetros (variables independientes *tto* y *años_c*)

Parámetro	B	Error típico	Interv. confianza de Wald 95%		Contraste de hipótesis			Exp(B)
			Inferior	Superior	Wald	gl	Sig.	
(Intersección)	,65	,13	,41	,90	26,67	1	,000	1,92
tto	-,63	,21	-1,05	-,22	8,86	1	,003	,53
años_c	,14	,03	,09	,19	28,43	1	,000	1,15
tto * años_c	,08	,04	,00	,17	3,46	1	,063	1,09
(Escala)	1							

- **Coefficiente $\hat{\beta}_0$.** El valor exponencial de la intersección (1,92) es el número estimado de recaídas para los pacientes que han recibido el tratamiento estándar (*tto* = 0) y llevan consumiendo 14 años (*años_c* = 0).
- **Coefficiente $\hat{\beta}_1$ (*tto*).** El coeficiente asociado a la variable *tto* recoge el efecto de esa variable cuando *años_c* = 0 (14 años de consumo). El signo negativo del coeficiente (-0,63) indica que el número de recaídas disminuye al aumentar la variable *tto*; por tanto, el número de recaídas es menor con tratamiento combinado (*tto* = 1) que con el estándar (*tto* = 0). El valor exponencial del coeficiente (0,53) indica que, entre los pacientes con 14 años de consumo, el número estimado de recaídas con el tratamiento combinado es un 53% del estimado con el tratamiento estándar.

- *Coeficiente $\hat{\beta}_2$ (*años_c*)*. El coeficiente asociado a la variable *años_c* recoge el efecto de esa variable cuando *tto* = 0 (tratamiento estándar). El valor exponencial del coeficiente (1,15) indica que, entre los pacientes que han recibido el tratamiento estándar, el número estimado de recaídas va aumentando un 15% con cada año más de consumo.
- *Coeficiente $\hat{\beta}_3$ (*tto* × *años_c*)*. El coeficiente de regresión asociado al efecto de la interacción (0,08) refleja cómo cambia la relación entre el *número de recaídas* y los *años de consumo* al cambiar de *tratamiento*. Su valor exponencial (1,09) indica que el coeficiente que relaciona el número de recaídas con los años de consumo es un 9% mayor con el tratamiento combinado que con el estándar. No obstante, esa diferencia no alcanza la significación estadística (*sig.* = 0,063).

Regresión de Poisson con tasas de respuesta

Aunque hasta ahora nos hemos centrado solamente en cómo modelar recuentos (número de eventos), la regresión de Poisson también permite modelar *tasas de respuesta*.

Una *tasa* es un número de eventos de algún tipo dividido por una línea base relevante. Por ejemplo, el número de recaídas dividido por el tiempo de seguimiento, o el número de accidentes de tráfico al año dividido por la cantidad de vehículos que circulan, o el número de cigarrillos/día dividido por el inverso del tiempo de exposición al tabaco, o el número de muertes que se producen al año dividido por el número de habitantes, etc.

Hasta ahora hemos estado asumiendo que los recuentos analizados se obtenían a partir de una línea base única o constante. En nuestro ejemplo sobre el número de recaídas de pacientes con problemas de alcoholismo hemos asumido que el registro del número de recaídas tras el tratamiento se obtenía observando a todos los pacientes un período de tiempo idéntico (dos años). Si todos los pacientes no hubieran sido observados durante el mismo periodo de tiempo (mismo número de meses) habría que reflejar este hecho de algún modo para poder incorporarlo al análisis. Lógicamente, no es lo mismo tener dos recaídas en, pongamos, 12 meses, que en 24 meses.

Por tanto, para trabajar con *tasas* es necesario crear dos variables en el archivo de datos: el *numerador* de la tasa (es decir, el recuento o número de eventos) y el *denominador* de la tasa (es decir, la línea base: el tiempo de seguimiento o exposición, el número de vehículos o habitantes, etc.). Al denominador de la tasa se le suele llamar **término de compensación** (*offset*).

La única diferencia entre modelar *recuentos* y modelar *tasas de respuesta* es que en el segundo caso es necesario incorporar al modelo el término de compensación. Esto significa que, en lugar de modelar $\log_e(\mu_i)$, se debe modelar $\log_e(\mu_i/\text{offset})$. Ahora bien, puesto que $\log_e(\mu_i/\text{offset}) = \log_e(\mu_i) - \log_e(\text{offset})$, el modelo de regresión de Poisson para tasas de respuesta queda de la siguiente manera:

$$\log_e(\mu_i) = \log_e(\text{offset}) + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \quad [7.16]$$

Debe repararse en el hecho de que el término de compensación de este modelo está en escala logarítmica.

En el archivo *Recaídas adicción alcohol*, la variable *seguimiento* recoge el número de meses de seguimiento que se ha hecho a cada paciente. Para incluir en el análisis el tiempo de seguimiento:

- En la pestaña **Tipo de modelo**, seleccionar la opción **Loglineal de Poisson**.
- En la pestaña **Respuesta**, trasladar la variable *recaídas* (número de recaídas) al cuadro **Variable dependiente**.
- En la pestaña **Predictores**, trasladar las variables *años_c* (años consumiendo, centrada) y *tto* (tratamiento) a la lista **Covariables** y la variable *log_seguimiento* (logaritmo de los meses de seguimiento) al cuadro **Variable de compensación**.
- En la pestaña **Modelo**, trasladar las variables *años_c* y *tto* a la lista **Modelo**.
- En la pestaña **Estadísticos**, marcar la opción **Incluir los valores exponenciales de las estimaciones de los parámetros**.

Aceptando estas selecciones, el SPSS ofrece, entre otros, los resultados que muestran las Tablas 7.16 y 7.17. La razón de verosimilitudes asociada al modelo que hemos propuesto para modelar la *tasa de recaídas* (93,19, ver Tabla 7.16) no es muy distinta de la asociada al modelo que hemos propuesto en el apartado anterior para modelar el *número de recaídas* (90,18, ver Tabla 7.10). Por tanto, aunque esto no tiene por qué ser así, el grado de ajuste de ambos modelos es muy parecido.

Sustituyendo los parámetros de la ecuación [7.16] por las estimaciones que ofrece la Tabla 7.17 se obtiene la siguiente ecuación de regresión:

$$\begin{aligned}\log_e(\hat{\mu}_{\text{recaídas}}) &= \log_e(\text{seguimiento}) + \hat{\beta}_0 + \hat{\beta}_1(\text{años_c}) + \hat{\beta}_2(\text{tto}) \\ &= \log_e(\text{seguimiento}) - 2,58 + 0,17(\text{años_c}) - 0,39(\text{tto})\end{aligned}\quad [7.17]$$

(debe tenerse en cuenta que el término *offset* no es un coeficiente de regresión, sino una variable en la que cada caso del archivo tiene su propia puntuación).

Tabla 7.16. Razón de verosimilitudes

Chi-cuadrado de la razón de verosimilitudes	gl	Sig.
93,19	2	,000

Tabla 7.17. Estimaciones de los parámetros (variables independientes *años_c* y *tto*)

Parámetro	B	Error típico	Interv. confianza de Wald 95%		Contraste de hipótesis			Exp(B)
			Inferior	Superior	Wald	gl	Sig.	
(Intersección)	-2,58	,12	-2,82	-2,34	435,45	1	,000	,08
años_c	,17	,02	,13	,21	63,88	1	,000	1,19
tto	-,39	,17	-,73	-,05	5,16	1	,023	,68
(Escala)	1,00							

Comparando el modelo propuesto para la *tasa de recaídas* (ecuación [7.17]) con el propuesto para el *número de recaídas* (ecuación [7.11]) se puede apreciar que únicamente la intersección muestra un cambio apreciable: ha pasado de 0,55 a -2,58. Los coeficientes asociados a las variables *años_c* y *tto* toman aproximadamente el mismo valor. Y ambos se interpretan en los términos ya conocidos. Lo único que diferencia a este modelo de *tasas* del modelo de *recuentos* es que en los pronósticos del modelo de *tasas* interviene el término de compensación.

Sobredispersión

El problema de la sobredispersión ya lo hemos tratado en el Capítulo 5 a propósito de la regresión logística binaria (ver el apartado *Dispersión proporcional a la media*). Para todo lo relativo al concepto de sobredispersión y a las consecuencias que se derivan de ella, lo dicho allí sirve también aquí; el concepto de sobredispersión sigue siendo el mismo y sus consecuencias también.

La media y la varianza de una distribución de Poisson son iguales (ver el Apéndice 1). Por tanto, para que la distribución de Poisson pueda representar apropiadamente el componente aleatorio del modelo propuesto, la varianza de los recuentos debe ser similar a su media.

Para cuantificar el grado de dispersión se suele utilizar un parámetro llamado **parámetro de escala**. Este parámetro de dispersión puede estimarse dividiendo la desvianza del modelo propuesto entre sus grados de libertad. Cuando la dispersión observada y la esperada son iguales, ese cociente toma un valor próximo a 1 (*equidispersión*). Un resultado mayor que 1 indica *sobredispersión*; valores mayores que 2 son problemáticos. Un resultado menor que 1 indica *infradispersión*; la infradispersión es infrecuente.

La desvianza, sus grados de libertad y el cociente entre ambos se ofrecen en la tabla de *estadísticos de bondad de ajuste*. La Tabla 7.18 muestra los estadísticos de bondad de ajuste correspondientes al modelo de regresión estimado en la Tabla 7.9, es decir, al modelo que incluye las variables independientes *años_c* y *tto*. La desvianza vale 101,97 y sus grados de libertad son 81 (el número de casos, 84, menos el número de coeficientes estimados, incluida la intersección). El cociente $101,97/81 = 1,26$ es la estimación que el procedimiento ofrece para el *parámetro de escala*. Se trata de un valor

Tabla 7.18. Estadísticos de bondad de ajuste

	Valor	gl	Valor / gl
Desvianza	101,97	81	1,26
Desvianza escalada	81,00	81	
Chi-cuadrado de Pearson	91,86	81	1,13
Chi-cuadrado de Pearson escalado	72,97	81	
Log verosimilitud	-129,86		
Criterio de información de Akaike (AIC)	265,72		
AIC corregido para muestras finitas (AICC)	266,02		
Criterio de información bayesiano (BIC)	273,01		
AIC consistente (CAIC)	276,01		

próximo a 1 que indica que el modelo propuesto no parece tener problemas con el grado de dispersión.

En el caso de que exista sobredispersión, su efectos indeseables pueden atenuarse aplicando una sencilla corrección a los errores típicos de los coeficientes. La corrección consiste en multiplicar cada error típico por la raíz cuadrada del valor estimado para el parámetro de escala (en nuestro ejemplo, por la raíz cuadrada de 1,26). Esta corrección hace aumentar el valor de los errores típicos y al aumentar el tamaño de los errores típicos disminuye el riesgo de declarar significativos efectos que no lo son. Las estimaciones de los coeficientes no cambian.

El procedimiento **Modelos lineales generalizados** ofrece la posibilidad de corregir automáticamente la dispersión observada aplicando bien una estimación del parámetro de escala basada en los datos (1,26 en nuestro ejemplo), bien un valor concreto fijado por el usuario. Estas opciones están disponibles en el menú desplegable **Método para el parámetro de escala** del subcuadro de diálogo correspondiente a la pestaña **Estimación**. Seleccionando la opción **Desviación** de ese menú desplegable se obtienen los estadísticos de bondad de ajuste de la Tabla 7.18.

Otra forma sencilla y bastante eficiente de atenuar los problemas derivados de la sobredispersión (también de la infradispersión) consiste en estimar los errores típicos de los coeficientes mediante algún método robusto. Para ello, basta con seleccionar, en la pestaña **Estimación**, la opción **Estimador robusto** del recuadro **Matriz de covarianzas**. Esta forma de estimar los errores típicos (conocida como método de Huber o método *sandwich*) no requiere que la distribución del componente aleatorio y la función de enlace estén correctamente especificadas.

Apéndice 7

Criterios de información

El procedimiento **Modelos lineales generalizados** ofrece varios estadísticos de bondad de ajuste. El estadístico de *Pearson* es el mismo que suele utilizarse para contrastar la hipótesis de bondad de ajuste con una variable y la hipótesis de independencia con dos variables:

$$X^2 = \sum_h (Y_h - \hat{\mu}_h)^2 / \hat{\mu}_h$$

(Y_h se refiere a los valores observados y $\hat{\mu}_h$ a los pronosticados; h se refiere a cualquier combinación de subíndices). El estadístico *desviación* es la razón de verosimilitudes que resulta de comparar la desviación del modelo propuesto y la del modelo saturado:

$$-2LL = -2 \left[\sum_h Y_h \log_e(Y_h / \hat{\mu}_h) \right]$$

Con muestras grandes, la distribución de estos dos estadísticos se aproxima a la distribución ji-cuadrado con un número de grados de libertad igual al número de casos menos el número de coeficientes de regresión estimados, incluida la intersección.

El *logaritmo de la verosimilitud* (LL) es la medida primaria de ajuste. Multiplicando LL por -2 se obtiene la *desvianza* ($-2LL$). El resto de criterios de información son modificaciones de $-2LL$ que penalizan (incrementando) su valor mediante, básicamente, alguna función del número de parámetros. AIC es el *criterio de información de Akaike* (Akaike, 1974):

$$AIC = -2LL + 2k$$

(k se refiere al número de coeficientes de regresión estimados, incluida la intersección). $AICC$ es el *criterio de información de Akaike corregido* (Hurvich y Tsai, 1989):

$$AICC = -2LL + 2k(k+1)/(n-k-1)$$

(n se refiere al tamaño muestral). BIC es el *criterio de información bayesiano* (Schwarz, 1978):

$$BIC = -2LL + k[\log_e(n)]$$

Y $CAIC$ es el *criterio de información de Akaike consistente* (Bozdogan, 1987):

$$CAIC = -2LL + k[\log_e(n) + 1].$$

La distribución binomial negativa y el problema de la sobredispersión

Ya hemos señalado que los problemas derivados de la presencia de sobredispersión pueden atenuarse multiplicando los errores típicos de los coeficientes de regresión por la raíz cuadrada del parámetro de escala. También hemos señalado que existe la posibilidad utilizar métodos robustos para estimar los errores típicos de los coeficientes.

Cuando la sobredispersión representa un problema realmente importante, una solución bastante eficaz consiste en sustituir la distribución de Poisson por la distribución **binomial negativa** (ver Gardner, Mulvey y Shaw, 1995). Esta distribución es muy parecida a la de Poisson, pero incluye un parámetro extra (el SPSS lo llama parámetro *auxiliar*) que permite que la media y la varianza de la distribución sean distintas, lo cual facilita la modelización de recuentos en presencia de sobredispersión. En una distribución de Poisson, la varianza es igual a la media: $\sigma_Y^2 = \mu$. En una distribución binomial negativa, $\sigma_Y^2 = \mu + \gamma\mu^2$. Si el parámetro γ vale cero, la distribución binomial negativa es idéntica a la de Poisson.

El procedimiento **Modelos lineales generalizados** permite contrastar la hipótesis nula de que el parámetro γ vale cero en la población. Para ello, tras seleccionar la variable *recaídas* como variable dependiente y las variables *años_c* y *tto* como covariables:

- En la pestaña **Tipo de modelo**, seleccionar la opción **Personalizado** y elegir la distribución **Binomial negativa** y la función de enlace **Logarítmica**. En el cuadro de texto **Valor**, introducir 0 como valor del parámetro auxiliar.
- En la pestaña **Estadísticos** marcar la opción **Contraste de multiplicadores de Lagrange para el parámetro de escala o para el parámetro auxiliar de la binomial negativa**.

Aceptando estas selecciones se obtienen, entre otros, los resultados que muestra la Tabla 7.19. El multiplicador de Lagrange permite contrastar la hipótesis nula de equidispersión ($\gamma = 0$). La

tabla ofrece tres niveles críticos, uno para cada posible hipótesis alternativa: *parámetro* < 0 se refiere a un contraste unilateral izquierdo (infradispersión), *parámetro* > 0 se refiere a un contraste unilateral derecho (sobredispersión) y *no direccional* se refiere a un contraste bilateral (varianza distinta de la media). Los resultados del ejemplo (*sig.* = 0,354 en contraste bilateral) indican que no parece haber problemas con la dispersión.

Tabla 7.19. Multiplicador de Lagrange (contraste sobre el parámetro auxiliar de la binomial negativa)

	Z	Significación observada (para cada hipótesis alternativa)		
		Parámetro < 0	Parámetro > 0	No direccional
Parámetro auxiliar	,93	,823	,177	,354

En el caso de que se rechace la hipótesis nula de que el parámetro γ vale cero, se puede intentar ajustar un modelo de regresión basado en la distribución binomial negativa. Pero, para esto, es necesario conocer o tener alguna idea acerca del valor del parámetro γ . Los resultados que se obtienen con esta estrategia (estadísticos de bondad de ajuste, coeficientes de regresión, etc.) se interpretan igual que cuando se utiliza la distribución de Poisson.

Modelos loglineales

El estudio de la relación entre variables categóricas lo hemos iniciado en el Capítulo 10 del primer volumen y lo hemos ampliado en el Capítulo 3 del segundo. Pero hasta ahora nos hemos limitado a estudiar una y dos variables. En este capítulo abordamos el estudio de múltiples variables categóricas mediante la aplicación de un tipo particular de modelos lineales llamados **logarítmico-lineales** o, abreviadamente, **loglineales**. Son modelos específicamente diseñados para estudiar la relación entre variables categóricas y, por tanto, especialmente útiles para analizar tablas de contingencias.

Ya sabemos que los modelos estadísticos más utilizados son los *lineales*. Son modelos en los que el valor esperado de un conjunto de observaciones (variable dependiente o respuesta) se interpreta como el resultado de una combinación lineal de varios efectos (variables independientes o predictoras). Con un modelo loglineal también se pretende explicar una variable dependiente a partir de una combinación lineal de variables independientes. Pero entre los modelos loglineales y el resto de modelos lineales estudiados en los capítulos anteriores existe una diferencia importante: la variable dependiente de un modelo loglineal no es ninguna de las variables incluidas en el análisis, sino la **frecuencia** con la que se repite cada patrón de variabilidad. El objetivo del análisis es encontrar la pauta de relación existente entre un conjunto de variables categóricas sin distinguir entre variables independientes y dependientes.

Para modelar frecuencias es preciso recurrir a alguna distribución de probabilidad que permita trabajar con números enteros no negativos. En los modelos loglineales se utiliza la distribución de Poisson.

Existen dos formas fundamentales de aproximación logarítmica al estudio de la relación entre variables categóricas: (1) los modelos **loglineales**, que sirven para estudiar la relación entre variables sin distinguir entre dependientes e independientes, y (2) los modelos **logit**, en los que una de las variables se considera dependiente.

Para profundizar en los contenidos de este capítulo puede consultarse Bishop, Fienberg y Holland (1975), Haberman (1978, 1979) o Powers y Xie (2000). Especialmente recomendables son, por su calidad y claridad, los trabajos de Agresti (2002, 2007) y Wickens (1989).

Tablas de contingencias

Cuando se trabaja con variables categóricas, los datos suelen organizarse en tablas de frecuencias de doble (triple, cuádruple, etc.) entrada en las que cada entrada representa un criterio de clasificación (una variable categórica). A estas tablas de frecuencias conjuntas las llamamos **tablas de contingencias**¹. Como resultado de la clasificación, las frecuencias (número, proporción o porcentaje de casos) aparecen organizadas en casillas que contienen información sobre la relación existente entre los criterios que conforman la tabla (ver Tabla 8.1).

Hasta ahora nos hemos limitado a estudiar tablas **bidimensionales** (tablas con dos variables o criterios de clasificación). En este capítulo empezaremos a estudiar tablas **multidimensionales** (más de dos variables).

Notación en tablas de contingencias

Para entender los modelos loglineales que estudiaremos en este capítulo es recomendable familiarizarse con la notación que utilizaremos para identificar cada elemento de una tabla de contingencias. Con las tablas bidimensionales utilizaremos la misma notación que en el Capítulo 3 del segundo volumen (ver Tabla 3.19). Con las tablas de más dimensiones seguiremos la misma lógica. Por ejemplo, en una tabla tridimensional, llamaremos X , Y y Z a las variables y les asignaremos subíndices i , j y k . Por tanto,

- $i = 1, 2, \dots, I$ (donde I es el número de categorías de la variable X).
- $j = 1, 2, \dots, J$ (donde J es el número de categorías de la variable Y).
- $k = 1, 2, \dots, K$ (donde K es el número de categorías de la variable Z).
- n_{ijk} = frecuencia observada en la casilla definida por la combinación de la categoría i de la variable X , la categoría j de la variable Y y la categoría k de la variable Z .

Para poder identificar cada casilla de una tabla tridimensional es necesario utilizar tres subíndices: uno por dimensión o variable. Los datos que ofrece la Tabla 8.1 pueden ayudarnos a familiarizarnos con la notación que utilizaremos en una tabla de contingencias y con la forma de obtener los diferentes totales marginales. Se trata de una tabla tridi-

¹ El término *contingencia* se refiere a la posibilidad de que algo ocurra. En una *tabla de contingencias* existen tantas posibilidades de que algo ocurra como combinaciones resultan de cruzar las categorías de las variables que definen la tabla. Por tanto, cada casilla de la tabla representa una posibilidad, es decir, una contingencia; de ahí que al conjunto de casillas de la tabla se le llame *tabla de contingencias*.

Tabla 8.1. Tabla de contingencias de *inteligencia* por *sexo* por *automensajes*

(X) Concepción inteligencia	(Y) Sexo	(Z) Tipo de automensajes			Totales de XY	Totales de X
		instrum.	atribuc.	otros		
destreza	hombres	21	7	4	32	41
	mujeres	3	4	2	9	
rasgo	hombres	5	10	3	18	59
	mujeres	6	28	7	41	
Totales de Z		35	49	16		100

mensional en la que una muestra de 100 sujetos se ha clasificado utilizando tres criterios (tres variables): *concepción que se tiene de la inteligencia* ($I = 2$; destreza, rasgo); *sexo* ($J = 2$; hombres, mujeres); y *tipo de automensajes* ($K = 3$; instrumentales, atribucionales, otros).

La frecuencia n_{123} se refiere a la primera categoría de la variable X ($i = 1$ = “destreza”), la segunda categoría de la variable Y ($j = 2$ = “mujeres”) y la tercera categoría de la variable Z ($k = 3$ = “otros”), lo cual nos sitúa en la casilla cuya frecuencia vale 2; por tanto, $n_{123} = 2$. Utilizando el mismo razonamiento se puede comprobar, por ejemplo, que n_{112} vale 7, y que n_{222} vale 28.

Para identificar los totales marginales, los subíndices i , j y k se sustituyen por el signo “+” allí donde es necesario. Un signo “+” como subíndice se refiere a todos los valores del subíndice al que sustituye. Así, por ejemplo, el total marginal n_{1++} es la suma de las frecuencias de la primera categoría de la variable X ($i = 1$ = “destreza”) en todas las categorías de las variables Y y Z ; por tanto, $n_{1++} = 21 + 3 + 7 + 4 + 4 + 2 = 41$. Y el total marginal n_{12+} es la suma de las frecuencias de la primera categoría de la variable X ($i = 1$ = “destreza”) y la segunda categoría de la variable Y ($j = 2$ = “mujeres”) en todas las categorías de la variable Z ; por tanto, $n_{12+} = 3 + 4 + 2 = 9$.

Para obtener una tabla de cuatro dimensiones basta con añadir una nueva variable (W por ejemplo) con su correspondiente subíndice (por ejemplo, l ; con $l = 1, 2, \dots, L$). Y cada elemento de la tabla queda identificado con cuatro subíndices.

Asociación en tablas de contingencias

Desde el punto de vista de la asociación entre variables, en una tabla de contingencias bidimensional solo cabe preguntarse si las dos variables que definen la tabla son independientes o están relacionadas, es decir, únicamente cabe la posibilidad de encontrar independencia o encontrar asociación (aunque ya sabemos que la asociación puede ser de diferentes tipos; ver Capítulo 3 del segundo volumen). Sin embargo, al añadir una nueva dimensión a la tabla y convertirla en tridimensional, las posibles pautas de asociación se incrementan de forma importante. Con tres variables (X , Y y Z) es posible

encontrar las siguientes pautas de asociación (por supuesto, cada pauta posee un significado concreto):

1. Las tres variables son independientes.
2. Existe asociación entre X e Y (pero no entre X y Z , ni entre Y y Z).
3. Existe asociación entre X y Z (pero no entre X e Y , ni entre Y y Z).
4. Existe asociación entre Y y Z (pero no entre X e Y , ni entre X y Z).
5. Existe asociación entre X e Y y entre X y Z (pero no entre Y y Z).
6. Existe asociación entre X e Y y entre Y y Z (pero no entre X y Z).
7. Existe asociación entre X y Z y entre Y y Z (pero no entre X e Y).
8. Existe asociación entre X e Y , entre X y Z y entre Y y Z .
9. Existe asociación entre X , Y y Z .

La pauta de asociación número 1 indica **independencia completa**. En ella solo están presentes los efectos principales de cada una de las variables, lo cual significa que las diferencias entre las frecuencias de las casillas únicamente reflejan diferencias en los totales marginales de cada variable individualmente considerada. Las tres variables de la tabla son mutuamente independientes o, si se prefiere, cada par de variables es condicionalmente independiente, dada la tercera. La independencia completa referida a una tabla tridimensional equivale a la hipótesis de independencia referida a una tabla bidimensional. Pero la diferencia entre ellas es importante. El rechazo de la hipótesis de independencia en una tabla bidimensional implica que las dos variables analizadas están relacionadas. En una tabla tridimensional solo indica que se da alguna de las restantes ocho pautas de asociación.

La pauta de asociación número 2 (al igual que la 3 y la 4) refleja **asociación parcial**. La presencia de la interacción XY y la ausencia de las interacciones XZ e YZ está indicando que las variables X e Y son condicionalmente dependientes, dada Z . Es decir, las variables X e Y están asociadas cualquiera que sea el nivel de Z que se considere (ver, en el Apéndice 3 del segundo volumen, el apartado *La paradoja de Simpson*).

La pauta de asociación número 5 (al igual que las pautas 6 y 7) es una pauta de **independencia condicional**. La ausencia de la interacción YZ indica que las variables Y y Z son condicionalmente independientes, dada X , es decir, Y y Z son independientes cualquiera que sea el nivel de la variable X que se considere.

La pauta de asociación número 8 indica que todas las variables están asociadas entre sí. A esta pauta de asociación parcial se le suele llamar **asociación homogénea**. La única interacción ausente es la de segundo orden: XYZ . Y esto significa que la relación entre cada par de variables es la misma en cada nivel de la tercera, es decir, la relación es la misma independientemente del nivel de la tercera variable que se considere.

Por último, la pauta de asociación número 9 es una pauta de **asociación completa**: están presentes todas las interacciones posibles. Indica que la asociación parcial entre cada par de variables cambia cuando cambia el nivel de la tercera variable.

Aunque en una tabla bidimensional solo cabe encontrar independencia o asociación, en una tabla tridimensional es posible encontrar nueve pautas de asociación distintas.

Y si se añade una cuarta variable, el número de pautas de asociación aumenta considerablemente. Pues bien, para poder estudiar estas múltiples pautas de asociación es necesario utilizar una estrategia que permita abordarlas de forma sistemática. Los **modelos loglineales** representan, quizá, la más útil de las herramientas disponibles para esto: no solo permiten realizar una aproximación sistemática al estudio de las pautas de asociación presentes en una tabla de contingencias, sino que, además, ofrecen la posibilidad de obtener estimaciones para los efectos que puedan resultar de interés.

Modelos loglineales jerárquicos

Para encontrar el modelo loglineal que mejor representa o describe la pauta de asociación presente en una tabla de contingencias aprenderemos a realizar cinco tareas:

1. Formular diferentes modelos loglineales.
2. Estimar las frecuencias esperadas que se derivan de un modelo loglineal.
3. Evaluar la calidad o ajuste de un modelo loglineal.
4. Seleccionar el modelo que podría dar cuenta de la pauta de asociación existente.
5. Analizar los residuos.

Cómo formular modelos loglineales

El modelo de independencia

Comencemos con el caso más simple: una tabla de contingencias **bidimensional**. Dos sucesos se consideran independientes cuando su probabilidad conjunta (su intersección) es igual al producto de sus probabilidades individuales (ver, en el Capítulo 2 del primer volumen, el apartado *Regla de la multiplicación*); es decir, dos sucesos A y B se consideran independientes cuando $P(A \cap B) = P(A)P(B)$. Aplicando esta misma regla a los sucesos *fila* y *columna* de una tabla de contingencias bidimensional, decimos que el suceso “fila = i ” es independiente del suceso “columna = j ” cuando la probabilidad de la intersección de ambos es igual al producto de sus probabilidades individuales:

$$P(\text{fila} = i \cap \text{columna} = j) = P(\text{fila} = i) P(\text{columna} = j) \quad [8.1]$$

es decir, cuando

$$\pi_{ij} = \frac{n_{i+}}{n} \frac{n_{+j}}{n} = \pi_{i+} \pi_{+j} \quad (\text{para todo } i \text{ y } j) \quad [8.2]$$

En consecuencia, si las filas son independientes de las columnas (es decir, si la variable X es independiente de la variable Y), la probabilidad de encontrar una observación en una casilla cualquiera es igual al producto de las probabilidades marginales de esa casilla.

Aplicando esta sencilla regla a las frecuencias de la Tabla 8.3 (ver más adelante), si las variables *sexo* y *tabaquismo* fueran independientes, la probabilidad del suceso *hombre-fumador* debería ser el resultado de multiplicar la probabilidad del suceso *hombre* (100/150) por la probabilidad del suceso *fumador* (60/150), es decir, 0,27.

Centrándonos en las frecuencias absolutas en lugar de hacerlo en las relativas, la *frecuencia esperada* (m_{ij}) de una casilla cualquiera, asumiendo que las filas son independientes de las columnas, puede obtenerse mediante

$$m_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j} \quad (\text{para todo } i \text{ y } j) \quad [8.3]$$

(llamaremos m_{ij} a las frecuencias esperadas, en lugar de μ_{ij} , para mantener la notación utilizada en los dos primeros volúmenes). Esta forma particular de interpretar los datos de una tabla de contingencias bidimensional consiste en hipotetizar que la frecuencia esperada de una casilla cualquiera es función directa de sus probabilidades marginales. Transformando [8.3] a escala logarítmica (por la comodidad de trabajar con un modelo aditivo en lugar de hacerlo con un modelo multiplicativo) obtenemos

$$\log_e(m_{ij}) = \log_e(n) + \log_e(\pi_{i+}) + \log_e(\pi_{+j}) \quad [8.4]$$

Por tanto, si se asume que las filas y las columnas de una tabla bidimensional son independientes, entonces el logaritmo de la frecuencia esperada de una casilla ij cualquiera es función lineal de los efectos de la i -ésima fila y de la j -ésima columna. Considerando que X es la variable que define las filas e Y la que define las columnas, y haciendo $\lambda = \sum_i \sum_j \log_e(m_{ij}) / (JK)$; $\lambda_{X(i)} = \sum_j \log_e(m_{ij}) / J - \lambda$; y $\lambda_{Y(j)} = \sum_i \log_e(m_{ij}) / I - \lambda$, es posible obtener por sustitución (ver, por ejemplo, Pardo y San Martín, 1994, pág. 557) una formulación alternativa de [8.4] similar a la que se utiliza en los modelos de análisis de varianza:

$$\log_e(m_{ij}) = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} \quad [8.5]$$

Después de transformadas todas las frecuencias de la tabla en sus correspondientes logaritmos, el término constante λ es la media de las frecuencias de toda la tabla; y los términos $\lambda_{X(i)}$ y $\lambda_{Y(j)}$ son las desviaciones de las medias marginales de cada fila y de cada columna respecto de la media total (siempre en escala logarítmica, que es la escala en la que se encuentran los términos lambda).

La ecuación [8.5] se conoce como **modelo loglineal de independencia** para una tabla de contingencias bidimensional. Se trata de un modelo de la familia de los modelos *lineales generalizados*: (1) los términos del componente sistemático se combinan aditivamente, (2) se asume que el componente aleatorio se ajusta, en cada casilla de la tabla, a una distribución de Poisson y (3) utiliza una función de enlace logarítmica.

El modelo de dependencia

También es posible formular un modelo loglineal para expresar la relación o dependencia entre X e Y . Para ello basta con introducir un término adicional referido a la interacción entre ambas variables. Considerando que la relación entre las variables tiene

que ver con las desviaciones que experimentan las frecuencias de las casillas respecto de sus correspondientes frecuencias marginales (de modo similar a como se hace con la interacción en un modelo de análisis de varianza), la relación XY puede definirse mediante

$$\lambda_{XY(ij)} = \log_e(m_{ij}) - \frac{\sum_j \log_e(m_{ij})}{J} - \frac{\sum_i \log_e(m_{ij})}{I} + \lambda \quad [8.6]$$

Por tanto, cuando no se asume que las filas son independientes de las columnas, las frecuencias esperadas de una tabla de contingencias bidimensional pueden expresarse completando [8.5] con [8.6]:

$$\log_e(m_{ij}) = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{XY(ij)} \quad [8.7]$$

La ecuación [8.7] se conoce como **modelo loglineal de dependencia** para una tabla de contingencias bidimensional. Y dado el parecido existente entre este modelo y el modelo de análisis de varianza de dos factores, es habitual utilizar una terminología similar a la del análisis de varianza para definir cada uno de sus componentes. Así,

- λ = media total del logaritmo de las frecuencias esperadas.
- $\lambda_{X(i)}$ = efecto de la i -ésima categoría de X (efecto de la i -ésima fila).
- $\lambda_{Y(j)}$ = efecto de la j -ésima categoría de Y (efecto de la j -ésima columna).
- $\lambda_{XY(ij)}$ = efecto de la interacción XY , es decir, de la combinación entre la i -ésima categoría de X y la j -ésima categoría de Y (o efecto de la combinación entre la i -ésima fila y la j -ésima columna).

Parámetros independientes

Las definiciones propuestas en el apartado anterior indican que los efectos principales se conciben, al igual que en un modelo de análisis de varianza, como desviaciones de las medias de las filas y de las columnas respecto de la media total. En consecuencia,

$$\sum_i \lambda_{X(i)} = \sum_j \lambda_{Y(j)} = 0 \quad [8.8]$$

Esto implica que existen $I - 1$ parámetros independientes $\lambda_{X(i)}$ asociados al efecto de las filas y $J - 1$ parámetros independientes $\lambda_{Y(j)}$ asociados al efecto de las columnas.

Del mismo modo (también a partir de las definiciones propuestas en el apartado anterior), el efecto de la interacción XY se concibe como la desviación de la media de cada casilla respecto de sus correspondientes medias marginales. En consecuencia,

$$\sum_i \lambda_{XY(ij)} = \sum_j \lambda_{XY(ij)} = 0 \quad [8.9]$$

Esto implica que, de los IJ parámetros $\lambda_{XY(ij)}$ asociados al efecto de la interacción XY , únicamente $(I - 1)(J - 1)$ son independientes. El número máximo de parámetros inde-

pendientes en un modelo loglineal es el número de casillas de la tabla, es decir, $I \times J$. Ese máximo se alcanza con el **modelo saturado**, que es el modelo que incluye todos los términos posibles (y que, según veremos, ofrece un ajuste perfecto). Pero el modelo saturado no solo no es el único disponible, sino que, puesto que incluye todos los términos posibles, tampoco suele ser el más interesante. La Tabla 8.2 muestra algunos modelos loglineales para una tabla de contingencias bidimensional.

Tabla 8.2. Modelos loglineales para una tabla de contingencias bidimensional

-
1. $\log_e(m_{ij}) = \lambda$
 2. $\log_e(m_{ij}) = \lambda + \lambda_{X(i)}$
 3. $\log_e(m_{ij}) = \lambda + \lambda_{X(i)} + \lambda_{Y(j)}$
 4. $\log_e(m_{ij}) = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{XY(ij)}$
-

El modelo 1 representa una situación de ausencia de efectos; asigna a todas las frecuencias esperadas el valor del único parámetro que incluye: λ (un mismo pronóstico para todas las frecuencias de la tabla).

El modelo 2 representa una situación en la que el único efecto presente es el de las filas ($\lambda_{X(i)}$). Asume que solo existe variabilidad entre las filas y que, por tanto, todas las categorías de la variable columna son igualmente probables. Incluye $1 + (I - 1)$ parámetros independientes: λ y tantos $\lambda_{X(i)}$ como filas menos una.

El modelo 3 incluye el efecto de las filas ($\lambda_{X(i)}$) y el de las columnas ($\lambda_{Y(j)}$). Por tanto, además del término constante, contiene $I - 1$ parámetros independientes para las filas y $J - 1$ para las columnas; en total, $1 + (I - 1) + (J - 1)$. Estos tres primeros modelos son modelos de *independencia*²: ninguno de ellos incluye un parámetro referido a la interacción filas-columnas.

El modelo 4 es el modelo *saturado*. Incluye los tres efectos posibles: el de las filas, el de las columnas y el de la interacción filas-columnas. Y contiene el máximo número posible de parámetros independientes: $1 + (I - 1) + (J - 1) + (I - 1)(J - 1) = IJ$, es decir, tantos parámetros independientes como casillas tiene la tabla. En una tabla de contingencias bidimensional, el modelo saturado es el único modelo de *dependencia*. Al incluir tantos parámetros como observaciones (casillas), sus predicciones son exactas (volveremos sobre esto).

Para entender mejor el significado de los parámetros de un modelo loglineal consideremos los datos de la Tabla 8.3. Se refieren a una muestra de 150 personas clasificadas aplicando dos criterios: *sexo* y *tabaquismo*. Las frecuencias observadas aparecen acompañadas, entre paréntesis, de sus logaritmos naturales.

² El tercero de ellos es el modelo de independencia completa. A los modelos que no incluyen términos referidos a todas las variables presentes en la tabla (como los modelos 1 y 2) se les llama *no comprensivos* (Bishop, Fienberg y Holland, 1975). Estos modelos carecen de significado, a no ser que uno de ellos demuestre ser el que mejor se ajusta a los datos, lo cual estaría indicando que alguna de las variables no contribuye a distinguir a unos sujetos de otros y, en consecuencia, que las dimensiones de la tabla deberían reducirse.

Tabla 8.3. Tabla de contingencias de sexo por *tabaquismo* (logaritmos entre paréntesis)

<i>Sexo</i>	<i>Tabaquismo</i>			<i>Medias</i>
	Fumadores	No fumadores	Exfumadores	
Hombres	30 (3,4012)	50 (3,9120)	20 (2,9956)	(3,4363)
Mujeres	30 (3,4012)	10 (2,3026)	10 (2,3026)	(2,6688)
<i>Medias</i>	(3,4012)	(3,1073)	(2,6492)	(3,0526)

En el modelo de dependencia [8.7], el logaritmo de cada frecuencia esperada se interpreta como una combinación lineal de *sexo*, *tabaquismo* y *sexo* $\times$ *tabaquismo*. Por ejemplo, el modelo loglineal de dependencia para la primera casilla de la tabla (es decir, para la casilla *hombres-fumadores*) adopta la forma:

$$\log_e(m_{11}) = \lambda + \lambda_{\text{sexo}(\text{hombres})} + \lambda_{\text{tabaco}(\text{fumadores})} + \lambda_{\text{sexo} \times \text{tabaco}(\text{hombres, fumadores})}$$

A partir de las definiciones propuestas para cada parámetro en [8.5] y [8.6] y sustituyendo cada frecuencia esperada m_h por su correspondiente observada n_h (donde h se refiere a cualquier combinación de subíndices) tal como se explica más adelante en el apartado *Cómo estimar las frecuencias esperadas de un modelo loglineal*, se obtiene:

$$\begin{aligned}\hat{\lambda} &= 3,0526 \\ \hat{\lambda}_{\text{sexo}(\text{hombres})} &= 3,4363 - 3,0526 = 0,3837 \\ \hat{\lambda}_{\text{tabaco}(\text{fumadores})} &= 3,4012 - 3,0526 = 0,3486 \\ \hat{\lambda}_{\text{sexo} \times \text{tabaco}(\text{hombres, fumadores})} &= 3,4012 - 3,4363 - 3,4012 + 3,0526 = -0,3837\end{aligned}$$

Por tanto, el modelo loglineal de dependencia ofrece, para la primera casilla de la Tabla 8.3 (es decir, para la casilla *hombres-fumadores*), el siguiente resultado:

$$\log_e(m_{11}) = 3,0526 + 0,3837 + 0,3486 - 0,3837 = 3,4012,$$

el cual coincide exactamente con el valor del logaritmo natural de la primera casilla de la tabla: $\log_e(n_{11}) = \log_e(30) = 3,4012$. Este resultado nos pone en la pista de algo que veremos enseguida: un modelo loglineal saturado (un modelo que incluye todos los términos posibles; en nuestro ejemplo, los términos que recogen los efectos de *sexo*, de *tabaquismo* y de *sexo* $\times$ *tabaquismo*) ofrece pronósticos perfectos, es decir, pronósticos que coinciden exactamente con las frecuencias observadas.

Tablas multidimensionales

Lo dicho para las tablas bidimensionales es fácilmente generalizable a tablas de más de dos dimensiones. En una tabla de contingencias *tridimensional*, por ejemplo, el modelo saturado adopta la siguiente forma:

$$\log_e(m_{ijk}) = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{Z(k)} + \lambda_{XY(ij)} + \lambda_{XZ(ik)} + \lambda_{YZ(jk)} + \lambda_{XYZ(ijk)} \quad [8.10]$$

El modelo saturado incluye todos los términos posibles: los relativos a los *efectos principales* de cada variable individualmente considerada, los relativos a las *interacciones de primer orden* entre cada par de variables y el relativo a la *interacción de segundo orden* entre las tres variables. Y, al igual que en los modelos para tablas de dos dimensiones, los parámetros siguen siendo desviaciones respecto de algún promedio relevante; consecuentemente,

$$\sum_i \lambda_{X(i)} = \sum_j \lambda_{Y(j)} = \sum_k \lambda_{Z(k)} = \sum_{ij} \lambda_{XY(ij)} = \cdots = \sum_{ijk} \lambda_{XYZ(ijk)} = 0 \quad [8.11]$$

A partir de aquí es fácil deducir que, en una tabla tridimensional, el número de parámetros independientes del modelo saturado es IJK (es decir, el número de casillas de la tabla). La Tabla 8.4 recoge esos parámetros desglosados para cada efecto.

Igualando a cero algunos de los términos del modelo saturado pueden obtenerse el resto de modelos loglineales disponibles para una tabla de contingencias tridimensional. No obstante, no todos ellos tienen la misma utilidad. Los *modelos jerárquicos* poseen algunas características que los hacen especialmente interesantes.

Tabla 8.4. Parámetros independientes en un modelo loglineal saturado

λ	:	1	} Total = IJK
$\lambda_{X(i)}$	:	$(I - 1)$	
$\lambda_{Y(j)}$	:	$(J - 1)$	
$\lambda_{Z(k)}$	:	$(K - 1)$	
$\lambda_{XY(ij)}$	:	$(I - 1)(J - 1)$	
$\lambda_{XZ(ik)}$	:	$(I - 1)(K - 1)$	
$\lambda_{YZ(jk)}$	:	$(J - 1)(K - 1)$	
$\lambda_{XYZ(ijk)}$	:	$(I - 1)(J - 1)(K - 1)$	

El principio de jerarquía

El número de modelos loglineales distintos que es posible formular en una tabla de contingencias aumenta considerablemente al añadir nuevas dimensiones a la tabla. Pero no todos los modelos que es posible formular resultan igualmente útiles. Exceptuando algunos modelos concretos que estudiaremos más adelante, los más utilizados son los **modelos jerárquicos**, que son modelos en los que *siempre que está presente un término de orden superior también lo están todos los términos de orden inferior que forman parte de él*. Por ejemplo, si un modelo incluye el término $\lambda_{XY(ij)}$, también debe incluir los dos términos contenidos en él, es decir, $\lambda_{X(i)}$ y $\lambda_{Y(j)}$. Así, por ejemplo, el modelo

$$\log_e(m_{ij}) = \lambda + \lambda_{X(i)} + \lambda_{XY(ij)} \quad [8.12]$$

no es un modelo jerárquico porque, estando presente el efecto de la interacción XY ($\lambda_{XY(ij)}$), no lo está el efecto de Y ($\lambda_{Y(j)}$). Un modelo jerárquico no permite interaccio-

nes entre dos variables si no está presente el efecto de cada variable por separado; ni interacciones triples si no están presentes las interacciones dobles que incluye; etc.

El principio de jerarquía permite utilizar sencillas abreviaturas o **símbolos** para identificar de forma rápida cada uno de los posibles modelos jerárquicos disponibles para una tabla de contingencias dada. Por ejemplo, el modelo saturado correspondiente a una tabla bidimensional puede identificarse mediante el símbolo $[XY]$, lo cual significa que se trata del modelo en el que está presente el término correspondiente a la interacción XY y, de acuerdo con el principio de jerarquía, todos los términos de orden inferior incluidos en esa interacción, es decir, los términos correspondientes a los efectos principales de X y de Y . La Tabla 8.5 muestra los modelos jerárquicos que es posible formular para una tabla de contingencias tridimensional. Cada modelo aparece acompañado de su correspondiente símbolo. A los elementos que forman parte del símbolo de un modelo jerárquico se les llama **configuraciones** (por ejemplo, el símbolo del modelo 2 viene definido por las configuraciones XY y Z ; el símbolo del modelo 3 viene definido por las configuraciones XY y XZ ; etc.). La Tabla 8.6 muestra el número de parámetros independientes asociados a los modelos de la Tabla 8.5.

Todas las consideraciones hechas sobre los modelos loglineales para tablas de contingencias tridimensionales son generalizables a tablas de cualquier número de dimensiones. Para una tabla dada, siempre existe un modelo saturado y un conjunto de modelos jerárquicos no saturados o restringidos que se obtienen igualando a cero algunos de los términos del modelo saturado.

Tabla 8.5. Algunos modelos loglineales jerárquicos para tablas de contingencias tridimensionales

<i>Modelo</i>	<i>Símbolo</i>
1. $\log_e(m_{ijk}) = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{Z(k)}$	$[X, Y, Z]$
2. $\log_e(m_{ijk}) = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{Z(k)} + \lambda_{XY(ij)}$	$[XY, Z]$
3. $\log_e(m_{ijk}) = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{Z(k)} + \lambda_{XY(ij)} + \lambda_{XZ(ik)}$	$[XY, XZ]$
4. $\log_e(m_{ijk}) = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{Z(k)} + \lambda_{XY(ij)} + \lambda_{XZ(ik)} + \lambda_{YZ(jk)}$	$[XY, XZ, YZ]$
5. $\log_e(m_{ijk}) = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{Z(k)} + \lambda_{XY(ij)} + \lambda_{XZ(ik)} + \lambda_{YZ(jk)} + \lambda_{XYZ(ijk)}$	$[XYZ]$

Tabla 8.6. Parámetros independientes asociados a los modelos loglineales de la tabla 8.5

<i>Modelo</i>	<i>Número de parámetros independientes</i>
1. $[X, Y, Z]$	$1 + (I - 1) + (J - 1) + (K - 1) = I + J + K - 2$
2. $[XY, Z]$	$1 + (I - 1) + (J - 1) + (K - 1) + (I - 1)(J - 1) = IJ + K - 1$
3. $[XY, XZ]$	$1 + (I - 1) + (J - 1) + (K - 1) + (I - 1)(J - 1) + (I - 1)(K - 1) = IJ + IK - I$
4. $[XY, XZ, YZ]$	$1 + (I - 1) + (J - 1) + (K - 1) + (I - 1)(J - 1) + (I - 1)(K - 1) + (J - 1)(K - 1) = IJ + IK + JK - I - J - K + 1$
5. $[XYZ]$	IKJ (ver Tabla 8.4)

Cómo estimar de las frecuencias esperadas de un modelo loglineal

Una vez elegido el modelo que previsiblemente servirá para explicar la variación observada en los datos, es necesario obtener las frecuencias esperadas que se derivan del mismo para, más tarde, poder valorar su ajuste.

Estimar las frecuencias esperadas de una tabla bidimensional es una tarea relativamente simple. En el modelo de independencia, las frecuencias esperadas $\{m_{ij}\}$ se definen mediante $m_{ij} = n\pi_{i+}\pi_{+j}$ (ver ecuación [8.3]). Y, cualquiera que sea el esquema de muestreo utilizado (ver Apéndice 8), las correspondientes estimaciones se obtienen sustituyendo los valores poblacionales por sus estimadores de máxima verosimilitud (ver, por ejemplo, Agresti, 1990, págs. 165-174): $\hat{m}_{ij} = n_{i+}n_{+j}/n$.

En el modelo saturado, $\hat{m}_{ij} = n_{ij}$. Sea cual sea el número de dimensiones de una tabla de contingencias, las frecuencias esperadas derivadas del modelo saturado siempre coinciden con las frecuencias observadas (esto es debido a que un modelo saturado incluye tantos parámetros independientes como casillas tiene la tabla).

En una tabla de contingencias de más de dos dimensiones es posible estimar las frecuencias esperadas con el método de máxima verosimilitud a través de ciertos estadísticos **mínimo-suficientes** (ver Goodman, 1970; Bishop, Fienberg y Holland, 1975). Un grupo de estadísticos es *suficiente* si permite reducir los datos de la tabla original y todavía es posible, con los datos restantes, efectuar las estimaciones. Con un estadístico mínimo-suficiente esa reducción de datos es máxima: permite ignorar la parte de la tabla que contiene información redundante para la estimación. En un modelo loglineal concreto, estos estadísticos mínimo-suficientes son las distribuciones marginales correspondientes a cada una de las configuraciones presentes en el símbolo del modelo³ (ver, en el Apéndice 8, el apartado *Estadísticos mínimo-suficientes*).

En algunos modelos loglineales, aunque las frecuencias esperadas siguen siendo función de los estadísticos mínimo-suficientes, no es posible estimarlas de forma directa a partir de los totales marginales de la tabla de frecuencias observadas. Esto es debido

³ Por ejemplo, en el modelo de independencia completa, $[X, Y, Z]$, los estadísticos mínimo-suficientes son n_{i++} , n_{+j+} y n_{++k} . Y, dado que las estimaciones máximo-verosímiles deben verificar que las configuraciones mínimo-suficientes de las frecuencias estimadas $\{\hat{m}\}$ sean iguales que las de las frecuencias observadas $\{n\}$, debe hacerse: $\hat{m}_{i++} = n_{i++}$, $\hat{m}_{+j+} = n_{+j+}$ y $\hat{m}_{++k} = n_{++k}$. Consecuentemente,

$$\hat{m}_{ijk} = n \hat{\pi}_{i++} \hat{\pi}_{+j+} \hat{\pi}_{++k} = n \frac{n_{i++}}{n} \frac{n_{+j+}}{n} \frac{n_{++k}}{n} = \frac{n_{i++} n_{+j+} n_{++k}}{n^2} \quad [8.13]$$

Del mismo modo, en un modelo de asociación parcial, $[XY, Z]$ por ejemplo, los estadísticos mínimo-suficientes son n_{ij+} y n_{++k} . Y las estimaciones de máxima verosimilitud de las frecuencias esperadas vendrán dadas por

$$\hat{m}_{ijk} = n \frac{n_{ij+}}{n} \frac{n_{++k}}{n} = \frac{n_{ij+} n_{++k}}{n} \quad [8.14]$$

Y en un modelo de independencia condicional, $[XY, XZ]$ por ejemplo, los estadísticos mínimo-suficientes son n_{ij+} y n_{i++} . Y las estimaciones máximo-verosímiles:

$$\hat{m}_{ijk} = n_{i++} \frac{n_{ij+}}{n_{i++}} \frac{n_{i+k}}{n_{i++}} = \frac{n_{ij+} n_{i+k}}{n_{i++}} \quad [8.15]$$

a que la función de verosimilitud no ofrece una solución única (ver Bishop, Fienberg y Holland, 1975, págs. 73-83). Tal es el caso, por ejemplo, del modelo [XY, XZ, YZ]. Cuando se da esta circunstancia, las frecuencias esperadas pueden obtenerse mediante métodos de cálculo iterativo. El procedimiento **Selección de modelo** del SPSS utiliza una versión del método de *ajuste proporcional iterativo* originalmente propuesto por Deming y Stephan (1940; ver Pardo 2002, págs. 88-89); el procedimiento **General** utiliza el algoritmo de Newton-Raphson (ver Haberman, 1974, 1978, 1979).

Estos métodos iterativos permiten realizar estimaciones cualquiera que sea el modelo loglineal y cualquiera que sea el número de dimensiones de la tabla de contingencias. Y las estimaciones que ofrecen coinciden con las estimaciones que se obtienen directamente a partir de las frecuencias marginales que se utilizan como estadísticos mínimo-suficientes (ver Bishop, Fienberg y Holland, 1975, págs. 85-87).

Una vez estimadas las frecuencias esperadas, ya es posible estimar los parámetros lambda que las han generado: puesto que los parámetros lambda son función únicamente de las frecuencias esperadas (ver ecuaciones [8.5] y [8.7]), una vez estimadas éstas, los parámetros lambda pueden estimarse simplemente sustituyendo en sus respectivas ecuaciones las frecuencias esperadas por sus estimaciones.

Cómo evaluar el ajuste o la calidad de un modelo loglineal

Tras formular el modelo y obtener las frecuencias esperadas que se derivan del mismo ya estamos en condiciones de comprobar si el modelo propuesto permite dar cuenta de los datos. Esto se hace valorando el grado en que las frecuencias observadas de la tabla se *ajustan* (se parecen) a las frecuencias esperadas que se derivan del modelo propuesto. Para valorar el ajuste de un modelo tenemos dos estadísticos diferentes⁴:

$$X^2 = \sum_h \frac{(n_h - \hat{m}_h)^2}{\hat{m}_h} \quad \text{y} \quad G^2 = 2 \sum_h n_h \log_e (n_h / \hat{m}_h) \quad [8.16]$$

El primero de ellos es el estadístico X^2 de Pearson (1911; Fisher, 1922). El estadístico G^2 es la *razón de verosimilitudes* (Fisher, 1924; Neyman y Pearson, 1928; Wilks, 1935; ver Rao, 1973). Estos estadísticos son asintóticamente equivalentes, es decir, ambos tienden a ofrecer el mismo resultado a medida que el tamaño muestral va aumentando. Con muestras grandes y bajo la hipótesis nula de que el modelo propuesto ofrece un buen ajuste a los datos, ambos estadísticos se aproximan a la distribución de probabilidad ji-cuadrado con gl grados de libertad.

El valor de gl es el resultado de restar al número de casillas de la tabla el número de parámetros independientes del modelo. En el Apéndice 8 se incluye un resumen de los grados de libertad asociados a cada uno de los modelos que es posible formular para tablas de contingencias de dos y tres dimensiones, así como el número de parámetros independientes asociados a cada modelo.

⁴ El subíndice h se refiere a todos los subíndices necesarios para identificar una casilla cualquiera. Así, en una tabla bidimensional, $h = ij$; en una tabla tridimensional, $h = ijk$; etc.

Siguiendo la estrategia habitual, deberá rechazarse la hipótesis nula de que el modelo se ajusta bien a los datos cuando el nivel crítico asociado a X^2 o G^2 sea menor que 0,05. En el caso de que no pueda rechazarse la hipótesis nula, podrá concluirse que el modelo propuesto ofrece un buen ajuste a los datos y, por tanto, que sirve para dar cuenta de la variabilidad observada.

En el contexto del ajuste de modelos loglineales, el estadístico G^2 es, en general, preferible al estadístico X^2 pues, como se explica en el siguiente apartado, G^2 posee una importante propiedad de descomposición que resulta especialmente útil a la hora de comparar modelos rivales.

Cómo seleccionar el mejor modelo loglineal

Al estudiar la relación entre variables categóricas es habitual encontrar que existe más de un modelo loglineal capaz de ofrecer un buen ajuste a los datos. Esto significa que, en una tabla de contingencias dada, la razón de verosimilitudes G^2 puede tomar un valor no significativo con más de un modelo concreto.

Si se tiene una hipótesis concreta, es decir, una idea previa acerca de la pauta de asociación estudiada, lo razonable será formular y ajustar el modelo que permita contrastar esa hipótesis. Pero tener hipótesis concretas es tanto más complicado cuanto mayor es el número de variables. En tablas de tres o más dimensiones existen tantas pautas de asociación (y modelos loglineales para representarlas) que no resulta nada fácil elegir una de ellas. En estos casos, y a falta de una hipótesis concreta que guíe la elección, es preferible proceder por pasos, añadiendo o quitando términos hasta encontrar el modelo capaz de describir la relación subyacente de la mejor manera posible. Y esto, sin perder de vista que el objetivo del análisis es encontrar el modelo que, además de tener algún significado teórico, guarde un buen equilibrio entre dos criterios que apuntan en direcciones opuestas: (1) ser lo bastante complejo como para ofrecer un buen ajuste a los datos (criterio de *máximo ajuste*) y, al mismo tiempo, (2) lo bastante simple como para ser fácilmente interpretable y lo más generalizable posible (criterio de *parsimonia*).

Fienberg (1970) y Goodman (1971; ver también Bonett y Bentler, 1983) han demostrado que es posible comparar dos modelos rivales restando sus respectivos valores G^2 (de idéntica manera a como se comparan modelos restando sus desvianzas; de hecho, el estadístico G^2 propuesto en [8.16] es la desviación). El único requisito para poder utilizar esta estrategia es que todos los términos de uno de los modelos estén incluidos en el otro, es decir, que se trate de modelos jerárquicos. Consideremos los siguientes modelos jerárquicos correspondientes a una tabla de contingencias tridimensional:

- a. [X,Y,Z]
- b. [XY,Z]
- c. [XY,XZ]
- d. [XY,XZ,YZ]
- e. [XYZ]

El orden en el que hemos presentado estos cinco modelos es tal que cada uno de ellos incluye todos los términos de los modelos que tiene por encima (ver Tabla 8.5). En este escenario, la razón de verosimilitudes G^2 posee dos importantes propiedades que no necesariamente se dan con el estadístico X^2 de Pearson:

$$\begin{aligned} 1. \quad G_a^2 &\geq G_b^2 \geq G_c^2 \geq G_d^2 \geq G_e^2 \\ 2. \quad G_a^2 &= (G_a^2 - G_b^2) + (G_b^2 - G_c^2) + (G_c^2 - G_d^2) + (G_d^2 - G_e^2) + G_e^2 \end{aligned} \quad [8.17]$$

La primera de estas dos propiedades afirma que el valor de G^2 va disminuyendo (o queda igual, aunque esto es improbable) a medida que se van incorporando nuevos términos al modelo; el límite de esta progresión se encuentra en el modelo saturado, cuya razón de verosimilitudes G^2 vale cero (el modelo saturado siempre ofrece pronósticos perfectos debido a que incluye el mayor número posible de términos; y si las frecuencias observadas y las esperadas son iguales, G^2 vale cero).

De la segunda propiedad se deduce que la disminución que se va produciendo en G^2 es consecuencia de los nuevos términos que se van incorporando al modelo. Por tanto, la diferencia entre, por ejemplo, los estadísticos G^2 correspondientes a los modelos c y d , es decir,

$$G_{c-d}^2 = G_c^2 - G_d^2 \quad [8.18]$$

representa la diferencia en el grado de ajuste de los modelos c y d . Y, puesto que se trata de la diferencia entre dos variables ji-cuadrado, G_{c-d}^2 también es una variable ji-cuadrado. Los grados de libertad de G_{c-d}^2 se obtienen restando los grados de libertad de los dos valores G^2 comparados. Por tanto, la ecuación [8.18] puede utilizarse para contrastar la hipótesis nula de que el término en el que difieren ambos modelos vale cero, es decir, $H_0: \lambda_{YZ(jk)} = 0$. El rechazo de esta hipótesis estaría indicando que el término extra que incluye el modelo d contribuye a reducir significativamente el desajuste del modelo c .

Cómo analizar los residuos

Al igual que ocurre con el resto de los modelos lineales, el análisis de los residuos no solo sirve para valorar la calidad del modelo propuesto sino para detectar posibles anomalías en los datos. Las estimaciones que se derivan de un modelo loglineal concreto suelen ser mejores en unas casillas que en otras y la constatación de este hecho puede arrojar luz sobre la pauta de asociación subyacente.

Ya sabemos que los **residuos** son las diferencias entre las frecuencias observadas y esperadas de cada casilla:

$$\hat{E}_h = n_h - \hat{m}_h \quad [8.19]$$

Cuanto mayores son los residuos (en valor absoluto), peor es el ajuste. El signo positivo o negativo de los residuos que más se alejan de cero puede estar indicando la presencia

de tendencias no bien representadas por el modelo. Una forma sencilla de evaluar estos residuos consiste en tipificarlos mediante

$$\hat{E}_{h(\text{tipificados})} = \hat{E}_h / \sqrt{\hat{m}_h} \quad [8.20]$$

Si no existen casillas con ceros estructurales (ver más adelante el apartado *Tablas incompletas*), estos **residuos tipificados** son componentes del estadístico X^2 de Pearson (si se suman tras ser elevados al cuadrado se obtiene el valor del estadístico X^2 ; por esta razón se les llama también *residuos de Pearson*). Con muestras grandes, la distribución de los residuos tipificados se aproxima a la normal con media cero y varianza igual a los grados de libertad del modelo divididos por el número de casillas de la tabla. La aproximación a la distribución normal es tanto mejor cuanto mayor es el tamaño muestral. El hecho de que la varianza de estos residuos no alcance el valor uno hace que no puedan ser interpretados exactamente como puntuaciones típicas.

Pierce y Schafer (1986) y McCullag y Nelder (1989) han definido otro tipo de residuos llamados **residuos de desviación** (*deviance residuals*). Se definen como la raíz cuadrada con signo de la contribución individual de cada casilla a la razón de verosimilitudes G^2 . Pueden calcularse fácilmente mediante:

$$\hat{E}_{h(\text{desviación})} = \text{sign}(\hat{E}_h) \sqrt{2 [n_h \log(n_h / \hat{m}_h) - (n_h - \hat{m}_h)]} \quad [8.21]$$

La distribución de estos residuos también se aproxima a la normal conforme el tamaño muestral va aumentando. Y, al igual que ocurre con los residuos tipificados respecto del estadístico X^2 de Pearson, los residuos de desviación poseen la importante propiedad de ser componentes del estadístico G^2 : cuando no existen ceros estructurales, la suma de estos residuos elevados al cuadrado coincide con la razón de verosimilitudes G^2 .

Haberman (1973) ha definido otro tipo de residuos tipificados muy utilizados llamados *tipificados corregidos* o, simplemente, **residuos corregidos**. A diferencia de lo que ocurre con los residuos tipificados de Pearson, los residuos corregidos sí se distribuyen $N(0, 1)$. Cualquiera que sea el modelo que se esté ajustando, estos residuos adoptan la forma⁵

$$\hat{E}_{h(\text{corregidos})} = \hat{E}_h / \hat{\sigma}_{\hat{E}_h} \quad [8.22]$$

⁵ En tablas bidimensionales, los residuos tipificados corregidos correspondientes al modelo de independencia $[X, Y]$ pueden obtenerse estimando el denominador de [8.22] mediante

$$\hat{\sigma}_{\hat{E}_h} = \sqrt{\hat{m}_{ij}(1 - n_{i\cdot}/n)(1 - n_{\cdot j}/n)} \quad [8.23]$$

El otro modelo de interés en tablas bidimensionales es el modelo de dependencia o saturado $[XY]$. Pero el análisis de los residuos asociados a un modelo saturado carece de sentido porque, simplemente, los residuos no existen (en un modelo saturado se verifica $\hat{E}_h = 0$ para todo h).

Las ecuaciones que permiten estimar $\hat{\sigma}_{\hat{E}_h}$ en tablas de tres o más dimensiones varían para cada modelo loglineal concreto. El lector interesado en conocer estas ecuaciones en los diferentes modelos disponibles para tablas tridimensionales puede consultar Haberman (1978, pág. 231). Y para un estudio detallado del procedimiento de cálculo de los residuos tipificados corregidos puede consultarse Haberman (1973; 1978, págs. 272-275).

Las distribuciones de los residuos de desviación y de los residuos corregidos se aproximan a la normal con media 0 y error típico 1. Y esto los hace fácilmente interpretables: con un nivel de confianza de 0,95, los residuos con valor absoluto mayor que 1,96 delatan casillas con más casos (si el residuo es positivo) o menos casos (si el residuo es negativo) de los que pronostica el modelo propuesto.

Cómo ajustar modelos loglineales jerárquicos con SPSS

El SPSS incluye tres opciones distintas para ajustar modelos loglineales: **General**, **Logit** y **Selección de modelo**. La opción **Selección de modelo** es la única que permite ajustar modelos jerárquicos por pasos; también permite ajustar un modelo concreto en un único paso, pero solo estima los parámetros del modelo saturado y no calcula los residuos tipificados corregidos ni los de desviación.

La opción **General** sirve para ajustar cualquier modelo loglineal (sea o no jerárquico) y, a diferencia de la opción **Selección de modelo**, permite obtener, entre otras cosas, los residuos tipificados corregidos y los de desviación, así como las estimaciones de los parámetros de cualquier modelo loglineal (y no solo del saturado). Puesto que la opción **General** no exige trabajar con modelos jerárquicos, resulta útil para ajustar modelos que representan hipótesis concretas (por ejemplo, cuasi-independencia, simetría, etc.). También permite incluir covariables en el análisis y analizar variables especiales como tasas de respuesta, tiempos de espera, etc. Al no exigir trabajar con modelos jerárquicos, con esta opción no es posible utilizar el ajuste por pasos.

La opción **Logit**, por último, permite ajustar modelos logit, que son un tipo particular de modelos loglineales en los que una de las variables categóricas se toma como variable dependiente o respuesta.

Para ajustar un modelo loglineal a los datos de una tabla de contingencias pueden seguirse dos estrategias alternativas. Si interesa contrastar una hipótesis particular, lo razonable es formular el modelo que represente la pauta de asociación correspondiente a esa hipótesis y ajustar ese modelo concreto (esto es preferible hacerlo con el procedimiento **General**; ver siguiente apartado). Si no se tiene una hipótesis concreta sobre la pauta de asociación existente entre las variables que conforman la tabla, es preferible comenzar con el modelo saturado e ir eliminando uno a uno términos no significativos hasta encontrar el modelo apropiado. Esta estrategia exige respetar el principio de jerarquía (pues la comparación entre los modelos de dos pasos sucesivos solo es posible si esos modelos son jerárquicos) y es, quizá, la de uso más generalizado pues permite detectar relaciones que de otro modo podrían pasar desapercibidas. Para ajustar **modelos loglineales jerárquicos** con el SPSS:

- Seleccionar la opción **Loglineal > Selección de modelo** del menú **Analizar** para acceder al cuadro de diálogo *Análisis loglineal: Selección de modelo*.

La opción **Utilizar eliminación hacia atrás** del recuadro **Construcción de modelos** permite ajustar modelos por pasos. En el ajuste por pasos, los términos del modelo de partida se van eliminando uno a uno comenzando por las interacciones de mayor orden (las que

incluyen un mayor número de variables). Este proceso continúa mientras quedan términos que no contribuyen significativamente al ajuste del modelo; por tanto, el proceso solo se detiene cuando eliminar cualquiera de los términos que permanecen en el modelo llevaría a una pérdida significativa de ajuste. Este proceso de eliminación hacia atrás se basa en el principio de jerarquía; en consecuencia, si un término de orden superior no puede ser eliminado del modelo, tampoco se eliminarán los términos de orden inferior contenidos en él. Para eliminar términos se utiliza la estrategia ya descrita en el apartado *Cómo seleccionar el mejor modelo loglineal*. Los términos se evalúan utilizando un nivel de significación de 0,05, pero el cuadro de texto **Probabilidad de eliminación** permite cambiar este valor. La opción **Introducir en un solo paso** está diseñada para evaluar el ajuste de un modelo loglineal concreto. Pero, puesto que esta estrategia no añade nada nuevo al ajuste por pasos, para ajustar un modelo concreto es preferible utilizar el procedimiento **General**, el cual incluye información adicional.

Si se utiliza la *eliminación hacia atrás*, el SPSS parte, por defecto, del modelo saturado; y si se opta por ajustar un modelo concreto *en un único paso*, el SPSS ofrece, por defecto, el ajuste del modelo saturado. El modelo saturado es, por tanto, el modelo de referencia tanto en la eliminación hacia atrás como en el ajuste en un único paso. Para utilizar como modelo de referencia un modelo distinto del saturado es necesario cambiar los valores por defecto de la opción **Modelo**. La opción **Saturado** del recuadro **Especificar un modelo** permite elegir el modelo saturado como punto de partida en la eliminación hacia atrás y como modelo de referencia en el ajuste de un modelo concreto. Es la opción que se encuentra activa por defecto. La opción **Personalizado** permite especificar modelos distintos del saturado. Para definir un modelo concreto es necesario seleccionar en la lista **Factores** las variables que se desea utilizar y trasladarlas a la lista **Clase generadora** utilizando el botón flecha y las opciones del menú desplegable del recuadro **Construir términos**. Para definir, por ejemplo, una interacción entre tres variables, hay que seleccionar esas tres variables en la lista **Factores**, la opción **Interacción** en el menú desplegable del recuadro **Construir términos** y pulsar el botón flecha.

Al construir un modelo personalizado debe tenerse en cuenta el principio de jerarquía. Esto significa que, en la lista **Clase generadora**, no hay que incluir los términos de menor orden incluidos en los de mayor orden ya definidos. Por ejemplo, si se incluye la interacción XY , no es necesario incluir (de hecho el cuadro de diálogo no lo permite) los efectos principales X e Y . Por tanto, la expresión *clase generadora* se refiere a las configuraciones que forman parte del símbolo de un modelo loglineal jerárquico.

Ajuste por pasos

En este apartado se explica cómo ajustar un modelo loglineal a los datos de la Tabla 8.1. Recordemos que la tabla contiene las frecuencias obtenidas al clasificar una muestra de 100 sujetos utilizando tres variables: *inteligencia* (concepción que se tiene de la inteligencia: destreza, rasgo), *sexo* (hombres, mujeres) y *automensajes* (tipo de mensajes autodirigidos al realizar una tarea de rendimiento: instrumentales, atribucionales y otros). El objetivo del análisis es encontrar la pauta de asociación existente entre estas

tres variables. Encontrar esa pauta de asociación equivale a encontrar el modelo loglineal capaz de ofrecer el mejor ajuste con el menor número de términos. Y la mejor forma de buscar ese modelo consiste en proceder por pasos comparando modelos alternativos que difieran en un solo término. En esta estrategia por pasos conviene comenzar con el modelo saturado (del que se sabe que $G^2 = 0$) e ir eliminando términos hasta llegar al modelo buscado. Para aplicar esta estrategia por pasos:

- Reproducir en el *Editor de datos* los datos de la Tabla 8.1 tal como muestra la Figura 8.1 o abrir el archivo *Loglineal jerárquico* que se encuentra en la página web del manual (se ha utilizado la función **Ponderar casos** del menú **Datos** para ponderar los casos con la variable *ncasos*).

Figura 8.1. Datos de la Tabla 8.1 reproducidos en el *Editor de datos*

	inteligencia	sexo	automensajes	ncasos
1	1	1	1	21
2	1	1	2	7
3	1	1	3	4
4	1	2	1	3
5	1	2	2	4
6	1	2	3	2
7	2	1	1	5
8	2	1	2	10
9	2	1	3	3
10	2	2	1	6
11	2	2	2	28
12	2	2	3	7

- Seleccionar la opción **Loglineal > Selección de modelo** del menú **Analizar** para acceder al cuadro de diálogo *Análisis loglineal: Selección de modelo* y trasladar las variables *inteligencia*, *sexo* y *automensajes* a la lista **Factores**.
- Manteniendo seleccionadas las variables *inteligencia* y *sexo* en la lista **Factores**, pulsar el botón **Definir rango** para acceder al subcuadro de diálogo *Análisis loglineal: Definir rango*. Introducir el código 1 en el cuadro de texto **Mínimo** y el código 2 en el cuadro de texto **Máximo** (estos códigos deben ser valores enteros; de todas las categorías que tenga una variable, se incluirán en el análisis las que se correspondan con los códigos *mínimo* y *máximo* más todas las comprendidas entre ellos; el resto de categorías quedarán fuera del análisis). Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Seleccionar la variable *automensajes* en la lista **Factores** y repetir la operación del párrafo anterior, pero utilizando los códigos 1 y 3 como valores **Mínimo** y **Máximo**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Opciones** para acceder al subcuadro de diálogo *Análisis loglineal: Opciones* y marcar las opciones **Estimaciones de los parámetros** y **Tablas de asocia-**

ción. Sustituir el valor 0,5 de la opción **Delta**⁶ por el valor 0. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor de resultados* ofrece la información que muestran las Tablas 8.7 a 8.15. La Tabla 8.7 indica, en el primer bloque (*casos*), que se están utilizando 12 casos no ponderados (*válidos*) que, en realidad, son 100 ponderados (*válidos ponderados*); también indica que no se ha desechado ningún caso por pertenecer a una categoría distinta de las incluidas en el análisis (*fuera de rango* = 0) y que no existen valores perdidos (*perdidos* = 0). El segundo bloque (*categorías*) recuerda con qué variables se va trabajar y el número de categorías de que consta cada una.

Tabla 8.7. Información sobre los datos

		N
Casos	Válido	12
	Fuera del rango	0
	Perdido	0
	Válido ponderado	100
Categorías	Inteligencia	2
	Sexo	2
	Automensajes	3

La Tabla 8.8 indica cuál es el modelo del que parte el análisis (*clase generadora* = *inteligencia* × *sexo* × *automensajes*), que no es otro que el modelo saturado, y algunos detalles relacionados con el proceso de estimación: el algoritmo de ajuste iterativo ha alcanzado el criterio de convergencia en la primera iteración, la diferencia más grande entre los marginales (estadísticos mínimo-suficientes) observados y estimados vale cero, y se ha utilizado un criterio de convergencia⁷ de 0,25.

Tabla 8.8. Información sobre la convergencia (modelo saturado)

Clase generadora	inteligencia*sexo*automensajes
Número de iteraciones	1
Diferencia máxima entre observados y marginales ajustados	,000
Criterio de convergencia	,250

A continuación de las Tablas 8.7 y 8.8, el SPSS ofrece otras dos tablas con las estimaciones y residuos del modelo saturado, y con los correspondientes estadísticos de ajuste. En estas tablas, puesto que el modelo de referencia es el saturado, las frecuencias espe-

⁶ El valor *delta* añade una constante a todas las frecuencias de la tabla para evitar los problemas derivados de la presencia de casillas vacías (esta constante afecta únicamente al modelo saturado). Puesto que en nuestro ejemplo no existen casillas vacías, no es necesario añadir ninguna constante a las frecuencias observadas.

⁷ El proceso de ajuste iterativo se detiene cuando la diferencia entre la estimación obtenida en un paso previo y la obtenida en el paso siguiente es menor que el valor de convergencia; este valor es, por defecto, 10^{-3} veces la frecuencia observada más grande o 0,25, el valor mayor de ambos. Este valor de convergencia puede cambiarse seleccionando cualquiera de las opciones del menú desplegable.

radas coinciden con las observadas y, consecuentemente, tanto los residuos como los estadísticos de ajuste valen cero (recordemos que el modelo saturado ofrece un ajuste perfecto). Se trata de información irrelevante (por conocida) que, no obstante, el SPSS se encarga de recordar.

La Tabla 8.9 ofrece los contrastes de los términos o efectos de orden K (K se refiere al número de variables que forman parte del efecto). La mitad superior muestra los contrastes para los *efectos de orden K o mayores*; puesto que se están utilizando tres variables, los efectos de orden $K = 3$ o mayores solo incluyen un efecto: el de orden 3 (la interacción entre las tres variables); los efectos de orden $K = 2$ o mayores incluyen los tres efectos de orden 2 (interacciones entre cada par de variables) y el efecto de orden 3; finalmente, los efectos de orden $K = 1$ o mayores incluyen los tres efectos de orden 1 (efectos principales), los tres de orden 2 y el efecto de orden 3.

La mitad inferior de la tabla muestra los contrastes para los *efectos de orden K* . Para cada efecto o grupo de efectos se ofrecen los grados de libertad (*gl*), el valor de los dos estadísticos de ajuste (la *razón de verosimilitudes* y el estadístico de *Pearson*, y el nivel crítico asociado a cada estadístico (*sig.*). La hipótesis nula que se contrasta en cada caso es que el efecto o grupo de efectos considerados valen cero (son nulos). Por tanto, estos contrastes permiten formarse una primera idea acerca de qué efectos estarán presentes en el modelo final: un efecto o grupo de efectos se considera significativo cuando se rechaza la hipótesis nula. Y, siguiendo la regla de decisión habitual en los contrastes de hipótesis, se rechaza la hipótesis nula cuando el nivel crítico (*sig.*) asociado a un efecto o grupo de efectos es menor que 0,05. En nuestro ejemplo, el resultado de estos contrastes indica, por ejemplo, que el término referido a la interacción triple ($K = 3$) no es significativo (*sig.* = 0,870); o que entre los términos referidos a las interacciones dobles o de primer orden ($K = 2$) existe al menos uno que es significativo (*sig.* < 0,0005); o que entre los términos referidos a los efectos principales ($K = 1$) existe al menos uno que es significativo (*sig.* < 0,0005).

Tabla 8.9. Contrastes de los efectos de orden k o mayores

	K	gl	Razón de verosimilitudes		Pearson	
			Chi-cuadrado	Sig.	Chi-cuadrado	Sig.
Efectos de orden K o mayores	1	11	66,69	,000	84,56	,000
	2	7	45,75	,000	54,64	,000
	3	2	,28	,870	,28	,871
Efectos de orden K	1	4	20,94	,000	29,92	,000
	2	5	45,47	,000	54,36	,000
	3	2	,28	,870	,28	,871

La Tabla 8.10 contiene una valoración de las *asociaciones parciales*, es decir, una valoración individual de cada término o efecto del modelo. La hipótesis nula que se contrasta en cada caso es que el correspondiente efecto vale cero (es decir, que es nulo). Los resultados de la tabla indican, por ejemplo, que entre las interacciones dobles, la única no significativa (la única que podría eliminarse del modelo sin pérdida de ajuste) es

la interacción $\text{sexo} \times \text{automensajes}$ ($\text{sig.} = 0,126$); y que entre los efectos principales solo es significativo el correspondiente a la variable automensajes ($\text{sig.} < 0,0005$). Los resultados de esta tabla permiten formarse una idea acerca de qué efectos estarán presentes en el modelo final; sin embargo, dado que las estimaciones de cada efecto dependen del modelo concreto que se está ajustando, lo razonable es completar el proceso por pasos para poder valorar qué efectos debe incluir el modelo final.

Tabla 8.10. Tabla de asociaciones parciales

Efecto	gl	Chi-cuadrado parcial	Sig.
inteligencia*sexo	1	13,50	,000
inteligencia*automensajes	2	9,04	,011
sexo*automensajes	2	4,14	,126
inteligencia	1	3,26	,071
sexo	1	,00	1,000
automensajes	2	17,68	,000

La Tabla 8.11 informa de los *parámetros independientes* del modelo saturado: estimaciones, errores típicos, valores tipificados (Z) e intervalos de confianza calculados al 95%. Dado que los parámetros asociados a un mismo efecto suman cero (ver [8.11]), no todos los parámetros son independientes. La tabla omite la información redundante. Así, por ejemplo, aunque el efecto principal de la variable sexo tiene asociados dos parámetros (uno por cada nivel de la variable: $\lambda_{\text{sexo}(\text{hombres})}$ y $\lambda_{\text{sexo}(\text{mujeres})}$), únicamente se estima el parámetro correspondiente a la primera categoría de la variable (en este caso, *hombres*). Por tanto,

$$\hat{\lambda}_{\text{sexo}(\text{hombres})} = 0,0950$$

Y, puesto que ambas estimaciones suman cero, el término $\hat{\lambda}_{\text{sexo}(\text{mujeres})}$ valdrá $-0,0950$. La tabla ofrece, para los efectos principales, las estimaciones correspondientes a todas las categorías excepto la última (que es redundante).

Por lo que se refiere a las interacciones dobles, la primera estimación corresponde a la combinación de la primera categoría de la primera variable y la primera categoría de la segunda variable; la segunda estimación corresponde a la combinación de la primera categoría de la primera variable y la segunda categoría de la segunda variable; etc. Para saber a qué efecto concreto corresponde cada parámetro hay que tener en cuenta que las categorías de la segunda variable rotan más rápido que las de la primera. Esto mismo vale también para las interacciones de mayor orden. Así, por ejemplo, aunque la interacción $\text{sexo} \times \text{automensajes}$ contiene seis parámetros (los resultantes de combinar las dos categorías de sexo con las tres de automensajes), la tabla solo ofrece las dos estimaciones no redundantes: la primera estimación corresponde a la combinación de las dos primeras categorías de ambas variables:

$$\hat{\lambda}_{\text{sexo} \times \text{automensajes}(\text{hombres, instrumentales})} = 0,3459$$

Y la segunda estimación corresponde a la combinación de la primera categoría de la primera variable (*sexo*) y la segunda categoría de la segunda variable (*automensajes*):

$$\hat{\lambda}_{\text{sexo} \times \text{automensajes (hombres, atribucionales)}} = -0,2125$$

Puesto que las estimaciones de las tres categorías de la variable *automensajes* suman cero en cada categoría de la variable *sexo*, el valor estimado para la tercera categoría de *automensajes* valdrá

$$\hat{\lambda}_{\text{sexo} \times \text{automensajes (hombres, otras)}} = -(0,3459 - 0,2125) = -0,1334$$

Y como las estimaciones correspondientes a las dos categorías de la variable *sexo* suman cero en cada categoría de la variable *automensajes*, las estimaciones de la segunda categoría de la variable *sexo* (mujeres) en cada categoría de la variable *automensajes* valdrán:

$$\hat{\lambda}_{\text{sexo} \times \text{automensajes (mujeres, instrumentales)}} = -0,3459$$

$$\hat{\lambda}_{\text{sexo} \times \text{automensajes (mujeres, atribucionales)}} = 0,2125$$

$$\hat{\lambda}_{\text{sexo} \times \text{automensajes (mujeres, otras)}} = 0,1334$$

Los valores tipificados (*Z*) se obtienen dividiendo cada estimación entre su error típico. Con tamaños muestrales grandes, la distribución de estos valores tipificados se aproxima a la normal con media 0 y desviación típica 1. Por tanto, pueden utilizarse para contrastar la hipótesis nula de que el correspondiente parámetro vale cero en la población. Se considera que un parámetro es significativamente distinto de cero cuando su valor tipificado tiene asociado un nivel crítico menor que 0,05 (o, lo que es lo mismo, cuando su valor absoluto es mayor que 1,96, que es el cuantil 97,5 en una distribución normal tipificada). Los intervalos de confianza permiten contrastar las mismas hipótesis nulas que los valores tipificados. Se considera que un parámetro es significativamente distinto de cero cuando su intervalo de confianza no incluye el valor cero. Así, por ejemplo, se

Tabla 8.11. Estimaciones de los parámetros (modelo saturado)

Efecto	Parám.	Estimación	Error típico	Z	Sig.	Intervalo de confianza al 95%	
						L. inferior	L. superior
inteligencia*sexo*automensajes	1	,0939	,18	,52	,605	-,26	,45
	2	-,0409	,17	-,24	,808	-,37	,29
inteligencia*sexo	1	,4382	,13	3,32	,001	,18	,70
inteligencia*automensajes	1	,3960	,18	2,18	,029	,04	,75
	2	-,3652	,17	-2,17	,030	-,70	-,03
sexo*automensajes	1	,3459	,18	1,91	,057	-,01	,70
	2	-,2125	,17	-1,26	,207	-,54	,12
inteligencia	1	-,2105	,13	-1,60	,110	-,47	,05
sexo	1	,0950	,13	,72	,471	-,16	,35
automensajes	1	,0831	,18	,46	,647	-,27	,44
	2	,4388	,17	2,60	,009	,11	,77

puede concluir que los dos parámetros independientes asociados al efecto de la interacción *inteligencia* $\times$ *automensajes* son distintos de cero, pues los correspondientes límites de confianza no incluyen el valor cero.

Una vez estimados los parámetros, la Tabla 8.12 ofrece un resumen de los resultados del proceso de *eliminación hacia atrás* partiendo del modelo saturado. En ese proceso se van contrastando *dos tipos de hipótesis nulas*. El primer tipo de hipótesis se refiere al *modelo* que se está ajustando en cada paso (*clase generadora*) y afirma que el modelo ofrece un buen ajuste a los datos. El segundo tipo de hipótesis se refiere a *efectos concretos* del modelo (*efecto eliminado*) y afirma que el efecto evaluado es nulo.

En el paso 0 se ajusta el modelo saturado (*clase generadora* = *inteligencia* $\times$ *sexo* $\times$ *automensajes*). Según se ha señalado ya, el modelo saturado se ajusta perfectamente a los datos; de ahí que el valor del estadístico de ajuste valga cero. Tras valorar el modelo saturado se ofrece un contraste del efecto que podría eliminarse en primer lugar en caso de ser nulo (*efecto eliminado* = *inteligencia* $\times$ *sexo* $\times$ *automensajes*). Se comienza valorando el efecto de mayor orden y la hipótesis nula que se contrasta es que ese efecto vale cero. El resultado del contraste indica que no puede rechazarse la hipótesis nula (*sig.* = 0,870); es decir, la interacción triple es no significativa y, consecuentemente, puede eliminarse del modelo sin pérdida de ajuste.

Al eliminar la interacción triple, el modelo resultante es el que contiene todas las interacciones dobles, es decir, el modelo de *asociación homogénea* (*clase generadora* = *inteligencia* $\times$ *sexo*, *inteligencia* $\times$ *automensajes*, *sexo* $\times$ *automensajes*). Este modelo es el que se evalúa en el paso 1. Puesto que la razón de verosimilitudes (*chi-cuadrado* = 0,28) tiene asociado un nivel crítico mayor que 0,05 (*sig.* = 0,870), se puede mantener la hipótesis nula y asumir que el modelo de asociación homogénea consigue un buen ajuste a los datos.

El siguiente paso del análisis consiste en averiguar si es posible eliminar alguno de los efectos que todavía permanecen en el modelo recién ajustado. Los efectos de mayor orden de este modelo son las tres interacciones dobles, de modo que el SPSS ofrece, todavía en el paso 1, un contraste de cada uno de esos efectos individualmente considerados. La hipótesis nula que se contrasta ahora es que el efecto evaluado vale cero. El nivel crítico asociado a cada contraste (*sig.*) indica que únicamente se mantiene la hipótesis nula referida al efecto de la interacción *sexo* $\times$ *automensajes* (*sig.* = 0,126). Y dado que esa interacción no parece contribuir al ajuste, lo razonable es prescindir de ella y, en el siguiente paso, ajustar el modelo [*inteligencia* $\times$ *sexo*, *inteligencia* $\times$ *automensajes*], que es un modelo de *independencia condicional*.

El ajuste de este modelo se ofrece en el paso 2. El nivel crítico (*sig.*) asociado a la razón de verosimilitudes vale 0,352, por lo que se puede asumir que el modelo ofrece un buen ajuste a los datos. Y puesto que este modelo no da problemas de ajuste, se debe continuar averiguando si es posible eliminar alguno de los efectos todavía incluidos en él. Los efectos de mayor orden ahora son las interacciones dobles *inteligencia* $\times$ *sexo* e *inteligencia* $\times$ *automensajes*, de modo que el SPSS ofrece un contraste de esos dos efectos individualmente considerados. La hipótesis nula que se contrasta en cada caso es que el efecto vale cero. Los niveles críticos asociados a estos dos efectos (*sig.* < 0,0005) indican que ambos son significativos: puesto que ambos poseen niveles críticos menores

que 0,05, en ambos casos se rechaza la hipótesis nula de que el efecto vale cero. Y esto significa que ninguno de los dos efectos debería quedar fuera del modelo: eliminar cualquiera de ellos llevaría a una pérdida significativa de ajuste.

En consecuencia, el modelo finalmente elegido es el que incluye esas dos interacciones dobles: [*inteligencia* × *sexo*, *inteligencia* × *automensajes*]. Y eso es justamente lo que se indica en el paso 3. Y, dado que se están ajustando modelos *jerárquicos*, el modelo final también incluye los tres efectos principales que contienen esas dos interacciones dobles. Expresando el modelo final en la notación propuesta para los modelos loglineales se obtiene

$$\log_e(m_{ijk}) = \lambda + \lambda_{\text{intelig}} + \lambda_{\text{sexo}} + \lambda_{\text{automen}} + \lambda_{\text{intelig} \times \text{sexo}} + \lambda_{\text{intelig} \times \text{automen}}$$

que es un modelo de *independencia condicional*: las variables *sexo* y *automensajes* son condicionalmente independientes, dada la variable *inteligencia*. Es decir, de las tres variables incluidas en el análisis, solo las variables *sexo* y *automensajes* son independientes entre sí, y lo son tanto entre los sujetos que conciben la inteligencia como un *rasgo* como entre los que la conciben como una *destreza*. De otro modo: las variables *inteligencia* y *sexo* están relacionadas y la forma de esta relación no cambia cuando cambia el tipo de *automensajes* que utilizan los sujetos; y las variables *inteligencia* y *automensajes* están relacionadas y la forma de esta relación es la misma entre los *hombres* y entre las *mujeres*.

Tabla 8.12. Pasos del proceso de eliminación hacia atrás

Paso	Efectos	Chi-cuadrado	gl	Sig.
0 Clase generadora	intelig*sexo*automen	,00	0	.
Efecto eliminado 1	intelig*sexo*automen	,28	2	,870
1 Clase generadora	intelig*sexo, intelig*automen, sexo*automen	,28	2	,870
Efecto eliminado 1	intelig*sexo	13,50	1	,000
2	intelig*automen	9,04	2	,011
3	sexo*automen	4,14	2	,126
2 Clase generadora	intelig*sexo, intelig*automen	4,42	4	,352
Efecto eliminado 1	intelig*sexo	22,89	1	,000
2	intelig*automen	18,44	2	,000
3 Clase generadora	intelig*sexo, intelig*automen	4,42	4	,352

Una vez identificado el modelo final, el SPSS ofrece información específica sobre él. En primer lugar informa, en una tabla idéntica a la 8.11, acerca de algunos detalles relativos a la convergencia del proceso de estimación (pero ahora, esa información se refiere al modelo de *independencia condicional* (*clase generadora* = *inteligencia* × *sexo*, *inteligencia* × *automensajes*).

Y a continuación ofrece las frecuencias y residuos (Tabla 8.13) y los estadísticos de ajuste (Tabla 8.14). La Tabla 8.13 contiene las frecuencias observadas y las esperadas (las derivadas del modelo final) en valor absoluto y porcentual, los residuos (diferencia entre las frecuencias observadas y las estimadas) y los residuos tipificados (re-

siduos de Pearson; ver [8.20]). Y la Tabla 8.14 ofrece el valor de los dos estadísticos de ajuste (la razón de verosimilitudes y el estadístico de Pearson), sus grados de libertad (*gl*) y el nivel crítico (*sig.*) asociado a cada uno de ellos.

Recordemos que los residuos de Pearson se distribuyen de forma aproximadamente normal con media cero y varianza (*gl*)/(*n*° de casillas). En el ejemplo, la varianza de los residuos vale $4/12 = 0,33$. Por tanto, el error típico de estos residuos (raíz cuadrada de la varianza) vale 0,58. Los residuos tipificados que se alejan más de dos errores típicos de cero están delatando casillas donde falla el ajuste. Los valores obtenidos indican que el ajuste es bueno en todas las casillas: el residuo tipificado más grande (1,02) se aleja de cero menos de dos errores típicos.

Tabla 8.13. Frecuencias y residuos (modelo final)

Intelig.	Sexo	Automensajes	Observado		Esperado		Residuos	Residuos tipificados
			Recuento	%	Recuento	%		
Destreza	Hombres	Instrumentales	21,00	21,0%	18,73	18,7%	2,27	,52
		Atribucionales	7,00	7,0%	8,59	8,6%	-1,59	-,54
		Otras	4,00	4,0%	4,68	4,7%	-,68	-,32
	Mujeres	Instrumentales	3,00	3,0%	5,27	5,3%	-2,27	-,99
		Atribucionales	4,00	4,0%	2,41	2,4%	1,59	1,02
		Otras	2,00	2,0%	1,32	1,3%	,68	,60
Rasgo	Hombres	Instrumentales	5,00	5,0%	3,36	3,4%	1,64	,90
		Atribucionales	10,00	10,0%	11,59	11,6%	-1,59	-,47
		Otras	3,00	3,0%	3,05	3,1%	-,05	-,03
	Mujeres	Instrumentales	6,00	6,0%	7,64	7,6%	-1,64	-,59
		Atribucionales	28,00	28,0%	26,41	26,4%	1,59	,31
		Otras	7,00	7,0%	6,95	6,9%	,05	,02

Tabla 8.14. Estadísticos de bondad de ajuste (modelo final)

	Chi-cuadrado	gl	Sig.
Razón de verosimilitudes	4,418	4	,352
Pearson	4,514	4	,341

Debe tenerse en cuenta que el procedimiento **Selección de modelo** no calcula los residuos de desviación (ecuación [8.21]) ni los tipificados corregidos (ecuación [8.22]), y que solo estima los parámetros del modelo saturado. Por tanto, una vez obtenido el modelo jerárquico que ofrece el mejor ajuste con el menor número de parámetros (modelo final), suele resultar bastante útil ajustar ese modelo mediante el procedimiento **General** para obtener toda esa información complementaria (ver más adelante el apartado *Modelos loglineales generales*).

Por último, dado que el modelo final incluye dos interacciones dobles (*inteligencia* $\times$ *sexo* e *inteligencia* $\times$ *automensajes*), se puede precisar el significado del modelo elegido analizando las tablas bidimensionales correspondientes a esas dos interacciones. Para ello:

- Seleccionar la opción **Estadísticos descriptivos > Tablas de contingencias** del menú **Analizar** para acceder al cuadro de diálogo *Tablas de contingencias* y trasladar la variable *inteligencia* a la lista **Filas** y las variables *sexo* y *automensajes* a la lista **Columnas**.
- Pulsar el botón **Casillas** para acceder al subcuadro de diálogo *Tablas de contingencias: Mostrar en las casillas* y marcar la opción **Tipificados corregidos** del recuadro **Residuos**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones el *Visor* ofrece los resultados que muestran las Tablas 8.15 y 8.16. Ambas incluyen los residuos corregidos calculados asumiendo que *inteligencia* y *sexo* son independientes. Estos residuos se distribuyen de forma aproximadamente normal, con media 0 y desviación típica 1; por tanto, los valores muy grandes en valor absoluto (mayores que 1,96 si se utiliza un nivel de confianza de 0,95) delatan casillas con más casos (residuo positivo) o menos (residuo negativo) de los que cabría esperar si las dos variables cruzadas fueran independientes. Consecuentemente, estos residuos pueden utilizarse para interpretar las pautas de asociación presentes en la tabla.

Los residuos corregidos de la Tabla 8.15 indican que, respecto de lo que cabría esperar si las variables *inteligencia* y *sexo* fueran independientes, entre los *hombres* se produce un desplazamiento significativo de casos desde la categoría *rasgo* (−4,7) hacia la categoría *destreza* (4,7), mientras que entre las mujeres se produce justamente la pauta contraria. Los residuos corregidos de la Tabla 8.16 indican que entre los sujetos que conciben la inteligencia como una *destreza* existe un desplazamiento significativo de casos desde la categoría *atribucionales* (−3,7) hacia la categoría *instrumentales* (4,1), mientras que entre los sujetos que conciben la inteligencia como un *rasgo* se observa justamente la pauta contraria.

Tabla 8.15. Tabla de contingencias de *inteligencia* por *sexo*

			Sexo		Total
			Hombres	Mujeres	
Inteligencia	Destreza	Recuento	32	9	41
		Residuos corregidos	4,7	-4,7	
	Rasgo	Recuento	18	41	59
		Residuos corregidos	-4,7	4,7	
Total		Recuento	50	50	100

Tabla 8.16. Tabla de contingencias de *inteligencia* por *automensajes*

			Automensajes			Total
			Instrumentales	Atribucionales	Otras	
Inteligencia	Destreza	Recuento	24	11	6	41
		Residuos corregidos	4,1	-3,7	-,3	
	Rasgo	Recuento	11	38	10	59
		Residuos corregidos	-4,1	3,7	,3	
Total		Recuento	35	49	16	100

Modelos loglineales generales

El procedimiento **Loglineal > General** que se describe en este apartado ofrece algunas ventajas sobre el procedimiento **Loglineal > Selección de modelo** estudiado en el apartado anterior. Con el procedimiento **General** es posible ajustar tanto modelos jerárquicos como no jerárquicos, obtener residuos tipificados corregidos y de desviación, estimar los parámetros de cualquier modelo loglineal, utilizar variables cuantitativas como covariables, etc. Y también es posible contrastar algunas hipótesis (simetría, cuasi-independencia, etc.) y utilizar algunas variables dependientes (tasas de respuesta, tiempos de espera, etc.) que pueden resultar especialmente interesantes en algunos contextos. Como contrapartida, puesto que el procedimiento **General** no se rige por el principio de jerarquía, no permite ajustar modelos por pasos. No obstante, incluso aunque el interés inicial del análisis se centre en la selección por pasos de un modelo loglineal jerárquico, una vez identificado el modelo idóneo con el procedimiento **Selección de modelo**, todavía puede resultar útil ajustar ese modelo con el procedimiento **General** para obtener la información adicional que ofrece.

Cómo ajustar un modelo concreto

En este apartado nos vamos a centrar en cómo obtener e interpretar la información que ofrece el procedimiento **Loglineal > General** por defecto. Y lo vamos a hacer ajustando el modelo de independencia condicional [*inteligencia* × *sexo*, *inteligencia* × *automensajes*] al que hemos llegado en el ejemplo del apartado anterior al analizar los datos de la Tabla 8.1 con una estrategia de ajuste por pasos. Para ajustar este modelo:

- Seleccionar la opción **Loglineal > General** del menú **Analizar** para acceder al cuadro de diálogo *Análisis loglineal general* y trasladar las variables *inteligencia*, *sexo* y *automensajes* a la lista **Factores**.
- Pulsar el botón **Modelo** para acceder al cuadro de diálogo *Análisis loglineal general: Modelo* y marcar la opción **Personalizado**⁸.
- Incluir en la lista **Términos del modelo** los términos *inteligencia*, *sexo*, *automensajes*, *inteligencia* × *sexo* e *inteligencia* × *automensajes*. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones el *Visor* ofrece los resultados que muestran las Tablas 8.17 a 8.20. La Tabla 8.17 señala que se están utilizando 12 casos válidos (que en realidad son 100 ponderados) y que no se ha desechado ningún caso por tener valor perdido. Respecto de la tabla de contingencias indica que posee 12 casillas, ninguna de las cuales

⁸ El procedimiento **Loglineal > General** ofrece, por defecto, el ajuste del modelo saturado. Para ajustar un modelo distinto del saturado deben seleccionarse las correspondientes variables en la lista **Factores y covariables** y trasladarlas a la lista **Términos del modelo** utilizando el botón flecha y las opciones del menú desplegable del recuadro **Construir términos**. Al seleccionar términos debe tenerse en cuenta que en el procedimiento **General** no rige el *principio de jerarquía*; por tanto, para definir un modelo concreto es necesario incluir todos sus términos.

tiene ceros estructurales o *a priori* ni ceros muestrales (ver más adelante el apartado *Tablas incompletas*). Y respecto de las variables incluidas en el análisis, menciona sus nombres (o etiquetas, si existen) y el número de categorías de cada una de ellas.

Tabla 8.17. Información sobre los datos

		N
Casos	Válidos	12
	Perdidos	0
	Válidos ponderados	100
Casillas	Casillas definidas	12
	Ceros estructurales	0
	Ceros de muestreo	0
Categorías	Inteligencia	2
	Sexo	2
	Automensajes	3

La Tabla 8.18 informa sobre algunos detalles del proceso de estimación: el número máximo de iteraciones se ha establecido en 20 (valor por defecto) y el criterio de convergencia o diferencia entre las estimaciones de dos iteraciones consecutivas (*tolerancia de la convergencia*) en 0,001. Se ha superado el criterio de convergencia en la iteración número 5: la diferencia mayor (absoluta y relativa) entre las estimaciones de las dos últimas iteraciones es menor que 0,001.

Tabla 8.18. Información sobre la convergencia

Número máximo de iteraciones	20
Tolerancia de convergencia	,00100
Máxima diferencia absoluta final	,00005
Máxima diferencia relativa final	,00007
Número de iteraciones	5

La Tabla 8.19 ofrece los dos estadísticos de bondad de ajuste: la razón de verosimilitudes y el estadístico de Pearson. Ambos aparecen acompañados de sus correspondientes grados de libertad (*gl*) y niveles críticos (*sig.*); puesto que en ambos casos el nivel crítico es mayor que 0,05, puede asumirse que el modelo propuesto ofrece un buen ajuste a los datos. En dos notas a pie de tabla se recuerda cuál es la distribución del componente aleatorio (*modelo: Poisson*) y qué términos concretos incluye el modelo loglineal que se está evaluando (*diseño*).

Tabla 8.19. Estadísticos de bondad de ajuste

	Valor ^{c,d}	gl	Sig.
Razón de verosimilitudes	4,42	4	,352
Chi-cuadrado de Pearson	4,51	4	,341

c. Modelo: Poisson

d. Diseño: constante + automensajes + inteligencia + sexo + inteligencia*sexo + inteligencia*automensajes

Por último, la Tabla 8.20 muestra, para cada una de las 12 casillas de la tabla, las frecuencias observadas (*observado*) y las esperadas (*esperado*), ambas en valor absoluto (*n*) y porcentual (%), los residuos en *bruto* o no tipificados, los residuos tipificados (ver ecuación [8.20]), los residuos tipificados corregidos (ver ecuación [8.22]), y los residuos de desviación (ver ecuación [8.21]).

Tabla 8.20. Frecuencias y residuos

Intelligen	Sexo	Automen	Observado		Esperado		Residuos	Residuos tipificados	Residuos corregidos	Residuos desviación
			n	%	n	%				
Destreza	Hombres	Instrum	21	21,0%	18,7	18,7%	2,268	,524	1,737	,514
		Atribuc	7	7,0%	8,6	8,6%	-1,585	-,541	-1,350	-,559
		Otras	4	4,0%	4,7	4,7%	-,683	-,316	-,729	-,324
	Mujeres	Instrum	3	3,0%	5,3	5,3%	-2,268	-,988	-1,737	-1,076
		Atribuc	4	4,0%	2,4	2,4%	1,585	1,020	1,350	,931
		Otras	2	2,0%	1,3	1,3%	,683	,595	,729	,552
Rasgo	Hombres	Instrum	5	5,0%	3,4	3,4%	1,644	,897	1,194	,836
		Atribuc	10	10,0%	11,6	11,6%	-1,593	-,468	-,941	-,479
		Otras	3	3,0%	3,1	3,1%	-,051	-,029	-,038	-,029
	Mujeres	Instrum	6	6,0%	7,6	7,6%	-1,644	-,595	-1,194	-,618
		Atribuc	28	28,0%	26,4	26,4%	1,593	,310	,941	,307
		Otras	7	7,0%	6,9	6,9%	,051	,019	,038	,019

Con tamaños muestrales grandes, tanto los residuos corregidos como los de desviación se distribuyen de forma aproximadamente normal con media igual a cero y desviación típica igual a uno (recordemos que los residuos de Pearson, aunque también se distribuyen de forma aproximadamente normal, tienen desviación típica menor que uno). Por tanto, cuando un modelo se ajusta bien a los datos, tanto los residuos corregidos como los de desviación deben tomar valores comprendidos entre $-1,96$ y $1,96$ (valores entre los que se encuentra el 95% de los casos en una distribución normal tipificada). En los resultados de la Tabla 8.20 se puede apreciar que todos los residuos tipificados corregidos y de desviación tienen valores comprendidos entre $-1,96$ y $1,96$. Por tanto, no parece que haya un problema de ajuste en ninguna de las casillas de la tabla.

El procedimiento también ofrece, por defecto, algunos gráficos con información útil. El primero de ellos contiene los tres diagramas de dispersión resultantes de combinar las frecuencias observadas, las esperadas y los residuos tipificados corregidos (ver Figura 8.2). Cuando un modelo se ajusta bien a los datos, la nube de puntos del diagrama correspondiente a las frecuencias observadas y a las esperadas muestra una pauta lineal; los puntos de este diagrama estarán tanto más en línea recta cuanto más se parezcan las frecuencias observadas y las esperadas (en el diagrama de nuestro ejemplo se observa una pauta claramente lineal). Por el contrario, los dos diagramas correspondientes a los residuos no deben seguir, idealmente, ningún tipo de pauta (en los diagramas de nuestro ejemplo no se observa ninguna pauta clara). El tamaño de los residuos debe ser independiente del tamaño de las frecuencias observadas; por tanto, la presencia

de alguna pauta de variación sistemática evidente podría estar indicando que la modelización loglineal no es apropiada para describir los datos.

Los otros dos gráficos que ofrece el procedimiento son diagramas de *probabilidad normal* (ver Figura 8.3). En el primero de ellos (izquierda) están representados los residuos tipificados corregidos (valor observado) y sus correspondientes valores esperados normales: si los residuos tipificados se distribuyen normalmente, los puntos del diagrama deben seguir una pauta lineal, es decir, deben estar alineados en torno a la diagonal trazada en el gráfico. El segundo de ellos (derecha) es un diagrama de probabilidad normal sin tendencias. En él están representadas las desviaciones de cada residuo respecto de su correspondiente valor esperado normal; es decir, las distancias verticales entre cada punto y la diagonal del gráfico de la izquierda. Si los residuos tipificados se distribuyen normalmente, el valor de esas desviaciones deben oscilar de forma aleatoria en torno al valor cero (representado por la línea horizontal). La presencia de pautas de variación no aleatorias (por ejemplo, pautas lineales o pautas curvilíneas) estaría indicando que la distribución de los residuos se aleja de la normalidad.

Figura 8.2. Diagramas de dispersión: frecuencias y residuos

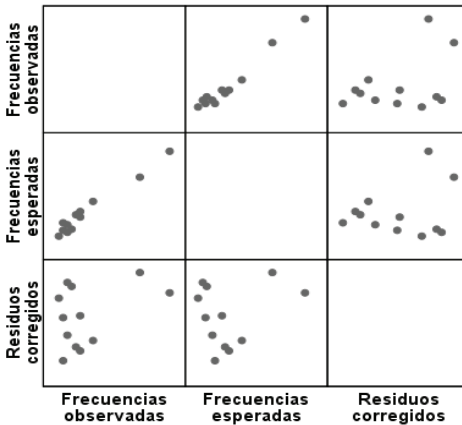
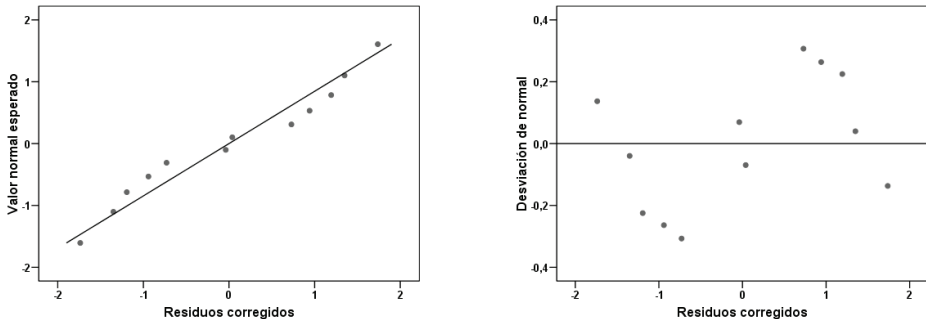


Figura 8.3. Diagramas de probabilidad normal (izqda.) y de probabilidad normal sin tendencias (dcha.)



En nuestro ejemplo, ambos gráficos muestran una pauta más o menos clara: los residuos negativos tienden a ser mayores que sus valores esperados normales y los residuos positivos tienden a ser menores que sus valores esperados normales. Sin embargo, esta pauta no es demasiado pronunciada; el eje vertical indica que los residuos observados se alejan no más de tres décimas de sus correspondientes esperados normales.

En el subcuadro de diálogo *Análisis loglineal general: Opciones* se pueden solicitar estos mismos gráficos para los residuos de desviación. Y el procedimiento **Selección de modelo** ofrece estos mismos gráficos para los residuos de Pearson.

Estimaciones de los parámetros

El procedimiento **Loglineal > General** permite estimar los parámetros de cualquier modelo loglineal. El recuadro **Mostrar** del subcuadro de diálogo *Análisis loglineal general: Opciones* contiene tres opciones que permiten controlar la información que se obtiene en relación con los parámetros del modelo. Estas opciones son **Matriz del diseño**, **Estimaciones** e **Historial de iteraciones**. La matriz del diseño contiene la información necesaria para saber qué casillas intervienen en el análisis y cuáles de ellas están involucradas en cada parámetro del modelo que se está ajustando. El historial de iteraciones muestra los valores que van tomando, en cada iteración, la razón de verosimilitudes y las estimaciones de los parámetros. Esta información tiene su interés, pero, por lo general, será suficiente con solicitar las estimaciones de los parámetros.

Las estimaciones que ofrece el procedimiento **General** se basan en una lógica distinta de la que utiliza el procedimiento **Selección de modelo**. Ya sabemos que cuando se trabaja con una variable categórica es necesario definir un esquema de codificación para poder analizar e interpretar su efecto. Un posible esquema de codificación consiste en comparar la frecuencia de cada categoría con el promedio de las frecuencias de todas ellas. Consideremos, por ejemplo, las tres categorías de la variable *automensajes*; para determinar si hay muchas o pocas respuestas *instrumentales* puede compararse la frecuencia de esa categoría con el promedio de las frecuencias de las tres categorías de la variable: *instrumentales*, *atribucionales* y *otras*. Así es como hemos definido en los apartados anteriores los parámetros de un modelo loglineal jerárquico (ver ecuaciones [8.5] y [8.6]) y así es también como define y estima los parámetros el procedimiento **Selección de modelo** ya estudiado (aunque solamente para el modelo saturado, pues no ofrece estimaciones para el resto de modelos).

Otro esquema de codificación consiste en comparar cada categoría con una de ellas que se toma como punto de referencia. Por ejemplo, para saber si hay muchas respuestas *instrumentales* puede compararse la frecuencia de esa categoría con la frecuencia de la categoría *otras*. Con este esquema de codificación, cualquier interpretación relacionada con la categoría *instrumentales* dependerá de la categoría de referencia elegida. Esta estrategia es la que utiliza el procedimiento **General**: fija en cero la última categoría de cada variable y estima los parámetros correspondientes al resto de categorías por comparación con esa categoría de referencia (ver Tabla 8.21). Por ejemplo, de los dos parámetros asociados a las dos categorías de la variable *inteligencia*, el último de ellos (el

correspondiente a la categoría *rasgo*) se fija en cero y se estima únicamente el correspondiente a la categoría *destreza*. De este modo, la categoría *rasgo* actúa como referente para interpretar el parámetro asociado a la categoría *destreza*.

Por tanto, los parámetros redundantes se fijan en cero (esta circunstancia se indica en una nota a pie de tabla) y solo se estiman los parámetros independientes o no redundantes. La tabla ofrece, para cada estimación, su error típico, su valor tipificado (Z) y los límites inferior y superior del intervalo de confianza calculado al 95%.

El valor tipificado de un parámetro (Z) se obtiene dividiendo su valor estimado entre su error típico. Con tamaños muestrales lo bastante grandes, estos valores tipificados se distribuyen normalmente con media 0 y desviación típica 1, por lo que pueden utilizarse para contrastar la hipótesis nula de que el correspondiente parámetro vale cero en la población. Un valor tipificado menor que -1,96 o mayor que 1,96 (cuantiles 2,5 y 97,5 de la distribución normal tipificada) debe llevar a rechazar la hipótesis nula de que el correspondiente parámetro vale cero.

De forma equivalente, se considera que un parámetro es significativamente distinto de cero y, por tanto, que el efecto o término que lo incluye debe estar presente en el modelo, cuando el valor cero no se encuentra dentro de los límites del correspondiente intervalo de confianza. En consecuencia, si se desea construir un modelo loglineal que ofrezca un buen ajuste a los datos de la Tabla 8.1, éste debe incluir además del término *constante*, los efectos principales *inteligencia*, *sexo* y *automensajes*, y las interacciones *inteligencia* × *sexo* e *inteligencia* × *automensajes*. A todos ellos les corresponde algún parámetro cuyo intervalo de confianza no incluye el valor cero.

Tabla 8.21. Estimaciones de los parámetros

Parámetro	Estimación	Error típ.	Z	Sig.	Intervalo de confianza al 95%	
					L. inferior	L. superior
Constante	1,94	,33	5,91	,000	1,30	2,58
[automensajes = 1]	,10	,44	,22	,827	-,76	,95
[automensajes = 2]	1,34	,36	3,76	,000	,64	2,03
[automensajes = 3]	,00 ^a	.	.	.	.	.
[inteligencia = 1]	-1,66	,60	-2,77	,006	-2,84	-,49
[inteligencia = 2]	,00 ^a	.	.	.	.	.
[sexo = 1]	-,82	,28	-2,91	,004	-1,38	-,27
[sexo = 2]	,00 ^a	.	.	.	.	.
[inteligencia = 1] * [automensajes = 1]	1,29	,63	2,04	,041	,05	2,53
[inteligencia = 1] * [automensajes = 2]	-,73	,62	-1,18	,239	-1,94	,49
[inteligencia = 1] * [automensajes = 3]	,00 ^a	.	.	.	.	.
[inteligencia = 2] * [automensajes = 1]	,00 ^a	.	.	.	.	.
[inteligencia = 2] * [automensajes = 2]	,00 ^a	.	.	.	.	.
[inteligencia = 2] * [automensajes = 3]	,00 ^a	.	.	.	.	.
[inteligencia = 1] * [sexo = 1]	2,09	,47	4,44	,000	1,17	3,02
[inteligencia = 1] * [sexo = 2]	,00 ^a	.	.	.	.	.
[inteligencia = 2] * [sexo = 1]	,00 ^a	.	.	.	.	.
[inteligencia = 2] * [sexo = 2]	,00 ^a	.	.	.	.	.

a. Este parámetro se ha definido como cero ya que es redundante.

Los parámetros de un modelo loglineal son función de las frecuencias esperadas (ver ecuaciones [8.5] y [8.6]). Pero las frecuencias esperadas también son función de los parámetros del modelo (ver ecuaciones [8.7] y [8.10]). Por tanto, las estimaciones de los parámetros pueden utilizarse para obtener las frecuencias que el modelo pronostica para cada casilla. Así, puesto que el modelo que se está ajustando es el de independencia condicional, el logaritmo de la frecuencia esperada de la primera casilla de la Tabla 8.1 (*destreza, hombre, instrumentales*) puede obtenerse mediante

$$\begin{aligned}\log_e(\hat{m}_{111}) &= \hat{\lambda} + \hat{\lambda}_{\text{instrum}} + \hat{\lambda}_{\text{destreza}} + \hat{\lambda}_{\text{hombres}} + \hat{\lambda}_{\text{instrum, destreza}} + \hat{\lambda}_{\text{hombres, destreza}} = \\ &= 1,939 + 0,095 - 1,663 - 0,823 - 1,291 + 2,092 = 2,931\end{aligned}$$

En consecuencia, $\hat{m}_{111} = e^{2,931} = 18,746$, que es justamente el valor que estima el procedimiento (salvando los ajustes del redondeo) para la frecuencia esperada asociada a esa primera casilla (ver Tabla 8.21).

Al solicitar las estimaciones de los parámetros, el procedimiento ofrece, además de las estimaciones de la Tabla 8.21, dos tablas adicionales con las correlaciones y las covarianzas entre las estimaciones. En términos generales, los parámetros de un modelo lineal (o loglineal) son linealmente independientes entre sí (de hecho, la independencia entre los parámetros es una característica fundamental de los modelos lineales). Por tanto, las correlaciones entre las estimaciones de los parámetros deben ser bajas. Correlaciones altas podrían estar indicando que el modelo loglineal propuesto no es el apropiado.

Estructura de las casillas

En un análisis loglineal convencional, todas las casillas de una tabla reciben, por defecto, un peso igual a uno. La opción **Estructura de las casillas** del cuadro de diálogo *Análisis loglineal general* permite alterar los pesos de las casillas. Los motivos por los que podría interesar modificar los pesos de las casillas son muy variados. En los siguientes apartados se explica cómo utilizar el procedimiento **Loglineal > General** para analizar algunas situaciones en las que es necesario alterar los pesos de las casillas. En concreto, se explica cómo analizar tablas que contienen *casillas con ceros estructurales*, cómo contrastar algunas hipótesis al analizar tablas cuadradas (tablas con el mismo número de filas y de columnas) y cómo analizar *tasas de respuesta*, como el número de accidentes dividido por el número de vehículos expuestos, o el número de muertes dividido por el número de pacientes, etc.

Tablas incompletas

La presencia de muchas casillas con frecuencias esperadas muy pequeñas (la *escasez* de datos) afecta negativamente tanto a la precisión de las estimaciones como al comportamiento de los estadísticos de ajuste (ver Agresti y Yang, 1987; Koehler, 1986; Koehler

y Larntz, 1980). Consecuentemente, en tablas con muchas casillas es importante utilizar muestras grandes para no tener que trabajar con frecuencias esperadas demasiado pequeñas (ver, en el Apéndice 9 del primer volumen, el apartado *Supuestos del estadístico X^2 de Pearson*).

Los problemas relacionados con la escasez de datos aumentan al trabajar con **tablas incompletas**, es decir, tablas con **casillas vacías** (casillas con frecuencia observada igual a cero). No obstante, no todas las casillas vacías tienen el mismo significado ni se tratan de la misma manera.

Al trabajar con tablas que contienen casillas vacías hay que distinguir entre (1) casillas con **ceros estructurales** o **a priori**, es decir, casillas que están vacías porque es imposible que pueda haber casos en ellas (por ejemplo, en un estudio clínico, el cruce de las variables *sexo* y *tipo de cáncer* arrojará inevitablemente algunas casillas vacías: hombre-útero, mujer-próstata, etc.) y (2) casillas con **ceros muestrales** o **a posteriori**, es decir, casillas en las que alguna frecuencia observada vale cero simplemente porque el tamaño de la muestra es demasiado pequeño en comparación con el número de casillas de la tabla y con la baja frecuencia con que aparecen ciertas combinaciones entre variables.

Ceros muestrales

Las casillas con ceros muestrales suelen aparecer cuando se utiliza gran cantidad de variables o variables con muchas categorías. Si la muestra es lo bastante grande, un cero muestral solo significa que la correspondiente combinación de categorías constituye un suceso raro. Y, por lo general, un pequeño porcentaje de casillas con ceros muestrales no representa un problema importante a no ser que los ceros muestrales generen un marginal vacío y ese marginal intervenga en el algoritmo de estimación (por decirlo de forma sencilla, si en un estudio sobre la opinión que las personas tienen sobre la eutanasia no se pregunta la opinión a las personas menores de 25 años, es evidente que no podrá concluirse nada sobre la opinión que tienen sobre la eutanasia las personas menores de 25 años).

No obstante, aunque las casillas con ceros muestrales no generen un marginal vacío, la presencia de casillas vacías tiene consecuencias poco deseables: las estimaciones se vuelven inestables (aumentan sus errores típicos) y los estadísticos de ajuste pierden precisión (la aproximación a la distribución ji-cuadrado se hace más lenta). Y como en tantas otras cuestiones relativas al tamaño muestral, no existe un criterio definitivo para decidir qué porcentaje de casillas vacías son admisibles para que el análisis funcione correctamente.

Con todo, los ceros muestrales pueden evitarse simplemente incrementando el tamaño de la muestra. Y, si esto no da resultado o no resulta fácil hacerlo, siempre existe la posibilidad, como propone Goodman (1971), de añadir una pequeña constante positiva a todas las frecuencias (0,5, por ejemplo) para eliminar los problemas computacionales derivados de la presencia de casillas vacías (el SPSS añade 0,5 puntos a cada casilla antes de estimar los parámetros de los modelos saturados).

Ceros estructurales

A diferencia de lo que ocurre con los ceros muestrales, los estructurales requieren un tratamiento especial⁹. Saber de antemano que en una casilla concreta no puede haber casos implica saber que la frecuencia esperada de esa casilla debe ser nula independientemente del modelo elegido.

Para entender lo que puede hacer un modelo loglineal con las casillas estructuralmente vacías, consideremos el caso de una tabla bidimensional $I \times J$ y llamemos C al conjunto de casillas no vacías: $C < IJ$. El análisis de una tabla bidimensional incompleta se realiza ajustando los mismos modelos loglineales ya descritos para tablas completas. La diferencia entre aplicar estos modelos a una tabla completa y aplicarlos a una tabla incompleta está únicamente en que, en presencia de casillas vacías, se verifica:

$$\sum_i \lambda_{X(i)} = \sum_j \lambda_{Y(j)} = \sum_i d_{ij} \lambda_{XY(ij)} = \sum_j d_{ij} \lambda_{XY(ij)} = 0 \quad [8.24]$$

con $d_{ij} = 1$ si $(ij) \in C$ y $d_{ij} = 0$ en cualquier otro caso. Las frecuencias esperadas de estos modelos se obtienen utilizando una modificación del método de estimación iterativo que asegura que las estimaciones obtenidas bajo un modelo particular valen cero en las casillas que contienen un cero *estructural* o *a priori*.

Una vez estimadas las frecuencias esperadas ya es posible evaluar el ajuste del modelo con el estadístico G^2 . Pero hay que tener en cuenta que los grados de libertad de una tabla de contingencias incompleta no son los mismos que los de su correspondiente tabla completa. En tablas incompletas, los grados de libertad se obtienen mediante:

$$gl = N_1 - N_2 - N_3, \quad [8.25]$$

donde: N_1 = “número de casillas de que consta la tabla”, N_2 = “número de parámetros independientes” y N_3 = “número de casillas con ceros estructurales”. La única complicación a la hora de determinar el valor de gl viene de N_2 . En una tabla incompleta, el número de parámetros que es necesario estimar (es decir, el número de parámetros independientes) es el mismo que en su correspondiente tabla completa, excepto por lo que se refiere a los parámetros relacionados con marginales vacíos (si existen), los cuales ya se sabe, *a priori*, que valen cero. En el siguiente apartado se muestra cómo utilizar el procedimiento **Loglineal > General** para analizar tablas con ceros estructurales.

Tablas cuadradas

Las tablas cuadradas son tablas bidimensionales con el mismo número de filas y de columnas. Por lo general, se construyen utilizando el mismo esquema de clasificación en las filas y en las columnas. En el ámbito de las ciencias sociales y de la salud no es

⁹ Existe abundante bibliografía relacionada con el análisis de *tablas incompletas*: Bishop y Fienberg (1969); Bishop, Fienberg y Holland (1975, págs. 177-210); Fienberg (1972; 1980, págs. 141-159); Goodman (1968); Haberman (1979, págs. 444-485); Mantel (1970), Wickens (1989, págs. 246-267); etc.

infrecuente encontrarse con la necesidad de analizar tablas cuadradas. Se obtienen, por ejemplo, cuando se clasifica una muestra de sujetos en una variable categórica en dos momentos distintos (como en los estudios de panel o en los diseños antes-después); o cuando se pide a una muestra de sujetos que clasifiquen por orden de importancia o preferencia dos objetos de una lista de k objetos; o cuando dos jueces o instrumentos de medida clasifican una muestra de sujetos en una misma variable categórica; etc.

La Tabla 8.22 muestra una tabla de este tipo con un grupo de 488 sujetos a los que se les ha pedido que seleccionen por orden de preferencia dos estímulos de una lista de cuatro (a, b, c, d). En las filas está representada la primera elección; en las columnas, la segunda. Obviamente, las casillas de la diagonal principal están vacías porque un sujeto no puede elegir el mismo estímulo dos veces; las casillas de la diagonal principal, por tanto, son casillas con ceros estructurales o *a priori*.

Aplicando el modelo loglineal de independencia a los datos de esta tabla se obtiene, para la razón de verosimilitudes, un valor de 341,44, el cual, con 9 grados de libertad, tiene asociado un nivel crítico menor que 0,0005. Este resultado indica que el modelo de independencia no ofrece un buen ajuste a los datos; y esto permite concluir que la primera y la segunda elección no son independientes.

Tabla 8.22. Tabla de contingencias de *primera elección* por *segunda elección*

<i>Primera elección</i>	<i>Segunda elección</i>				<i>Totales</i>
	1 = a	2 = b	3 = c	4 = d	
1 = a	0	19	28	14	61
2 = b	14	0	89	42	145
3 = c	23	92	0	66	181
4 = d	15	38	48	0	101
<i>Totales</i>	52	149	165	122	488

La razón por la cual el modelo de independencia no consigue un buen ajuste a los datos de la Tabla 8.22 hay que buscarla en las casillas vacías en la diagonal principal (los residuos tipificados corregidos más grandes en valor absoluto se dan en esa diagonal). Si se ignoran estas casillas, cabe la posibilidad de que el estímulo elegido en segundo lugar sea independiente del elegido en primer lugar. Para valorar esta circunstancia puede ajustarse un modelo loglineal de independencia forzando que las estimaciones de las frecuencias esperadas de la diagonal principal valgan cero.

Cuasi-independencia

A la hipótesis de independencia referida a la parte de la tabla que no contiene ceros estructurales se le llama hipótesis de **cuasi-independencia**. Y es posible formular modelos loglineales para contrastar esta hipótesis cualquiera que sea la ubicación de las

casillas con ceros estructurales. Por ejemplo, el modelo loglineal que permite poner a prueba la hipótesis de cuasi-independencia excluyendo del análisis las casillas de la diagonal principal adopta la siguiente forma:

$$\log_e(m_{ij}) = \lambda + \lambda_{x(i)} + \lambda_{y(j)} + \delta_i I \quad (I = 1 \text{ si } i = j; I = 0 \text{ si } i \neq j) \quad [8.26]$$

El término δ_i combinado con la variable indicador I es el que permite tratar por separado las casillas de la diagonal principal. Puesto que δ_i vale cero en todas las casillas excepto en las de la diagonal principal ($i = j$), en la estimación de los I parámetros δ_i únicamente intervienen las casillas de esa diagonal.

La hipótesis de cuasi-independencia no solo sirve para estudiar la asociación entre dos variables cuando se desea excluir del análisis las casillas que contienen ceros estructurales. También sirve para contrastar la hipótesis de independencia cuando, no estando vacías las casillas de la diagonal principal (o de cualquier otra parte de la tabla), no se desea que la información que contienen esas casillas forme parte del análisis.

Por ejemplo, en un estudio sobre movilidad social, al cruzar las variables *zona de residencia en 1990* y *zona de residencia en 2010*, dado que la mayoría de las personas no suelen cambiar de zona de residencia, cabe esperar que sea justamente en las casillas de la diagonal principal donde se concentre el mayor número de casos. El análisis de una tabla de este tipo mediante el modelo loglineal de independencia llevaría a la conclusión de que las variables estudiadas no son independientes justamente por la acumulación de casos en la diagonal principal. En estos casos, el modelo de cuasi-independencia, precisamente porque permitiría estudiar la asociación entre ambas variables prescindiendo de la diagonal principal, podría utilizarse para averiguar si las personas de una determinada zona tienden o no a desplazarse a otra determinada zona.

Para ajustar un modelo loglineal de cuasi-independencia con el procedimiento **Loglineal > General** es necesario crear una variable adicional cuyos valores indiquen qué casillas son las que contienen ceros estructurales (o qué casillas se desea dejar fuera del análisis). La Figura 8.4 muestra cómo reproducir en el *Editor de datos* las frecuencias de la Tabla 8.22. Hemos creado las tres variables necesarias para reproducir los datos de la tabla (*primera* = “primera elección”, *segunda* = “segunda elección” y *ncasos*) más una variable adicional (*casillas*) para indicar a qué combinaciones entre niveles les corresponde una casilla válida (*casillas* = 1) o una casilla con cero estructural (*casillas* = 0). Para ajustar el modelo de cuasi-independencia a los datos de la Tabla 8.22:

- Reproducir los datos de la Tabla 8.22 tal como muestra la Figura 8.4 y ponderar el archivo con la variable *ncasos* utilizando la opción **Ponderar casos** del menú **Datos** (o descargar el archivo *Loglineal cuasi-independencia* de la página web del manual).
- En el cuadro de diálogo *Análisis loglineal general*, trasladar las variables *primera* y *segunda* a la lista **Factores** y la variable *casillas* al cuadro **Estructura de las casillas**.
- Pulsar el botón **Modelo** para acceder al subcuadro de diálogo *Análisis loglineal general: Modelo*, marcar la opción **Personalizado** y definir, como **Términos del modelo**, los dos efectos principales *primera* y *segunda*. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, se obtienen, entre otros, los resultados que muestran las Tablas 8.23 y 8.24. La primera de ellas ofrece las frecuencias observadas, las esperadas y varios tipos de residuos. Puede comprobarse en la tabla que las casillas con ceros

Figura 8.4. Datos de la Tabla 8.22 reproducidos en el *Editor de datos*

	primera	segunda	ncasos	casillas
1	a	a	0	0
2	a	b	19	1
3	a	c	28	1
4	a	d	14	1
5	b	a	14	1
6	b	b	0	0
7	b	c	89	1
8	b	d	42	1
9	c	a	23	1
10	c	b	92	1
11	c	c	0	0
12	c	d	66	1
13	d	a	15	1
14	d	b	38	1
15	d	c	48	1
16	d	d	0	0

Tabla 8.23. Frecuencias y residuos (cuasi-independencia)

Estímulo primera elección	Estímulo segunda elección	Observado		Esperado		Resid. Resid.	Resid. tipificad.	Resid. corregid.	Resid. desvian.
		n	%	n	%				
a	a	0	,0%	,00	,0%	.	.	.	.
	b	19	3,9%	19,70	4,0%	-,70	-,16	-,21	-,16
	c	28	5,7%	27,35	5,6%	,65	,12	,19	,12
	d	14	2,9%	13,95	2,9%	,05	,01	,02	,01
b	a	14	2,9%	16,70	3,4%	-2,70	-,66	-,88	-,68
	b	0	,0%	,00	,0%	.	.	.	.
	c	89	18,2%	84,96	17,4%	4,04	,44	1,00	,44
	d	42	8,6%	43,34	8,9%	-1,34	-,20	-,34	-,20
c	a	23	4,7%	24,94	5,1%	-1,94	-,39	-,60	-,39
	b	92	18,9%	91,35	18,7%	,65	,07	,16	,07
	c	0	,0%	,00	,0%	.	.	.	.
	d	66	13,5%	64,71	13,3%	1,29	,16	,32	,16
d	a	15	3,1%	10,36	2,1%	4,64	1,44	1,72	1,35
	b	38	7,8%	37,95	7,8%	,05	,01	,01	,01
	c	48	9,8%	52,69	10,8%	-4,69	-,65	-1,21	-,66
	d	0	,0%	,00	,0%	.	.	.	.

estructurales (las casillas de la diagonal principal), están correctamente estimadas: todas las frecuencias esperadas correspondientes a esas casillas valen cero. Y el nivel crítico asociado a la razón de verosimilitudes vale 0,673 (ver Tabla 8.24), lo cual indica que el modelo de cuasi-independencia ofrece un buen ajuste a los datos. Por tanto, si se excluyen del análisis las casillas de la diagonal principal, no parece que el estímulo que se elige en primer lugar esté relacionado con el que se elige en segundo lugar. Los residuos corregidos y de desviación (todos ellos toman valores comprendidos entre -1,96 y 1,96) confirman que el modelo de cuasi-independencia se ajusta bien a los datos.

Tabla 8.24. Estadísticos de bondad de ajuste (cuasi-independencia)

	Valor	gl	Sig.
Razón de verosimilitudes	3,17	5	,673
Chi-cuadrado de Pearson	3,39	5	,640

Simetría completa

En las tablas de contingencias cuadradas es posible contrastar otras hipótesis además de la de independencia y cuasi-independencia. Una de estas otras hipótesis es la de **simetría completa** o *absoluta* o, simplemente, **simetría** (se trata de la misma hipótesis que se contrasta con las pruebas de McNemar y Bowker (ver Capítulo 3 del segundo volumen).

Una tabla de contingencias 2×2 es simétrica cuando las dos probabilidades que se encuentran fuera de la diagonal principal son iguales, es decir, cuando $\pi_{12} = \pi_{21}$. Una tabla $I \times J$ es simétrica cuando las probabilidades de las casillas simétricamente opuestas respecto de la diagonal principal son iguales; es decir, cuando $\pi_{ij} = \pi_{ji}$.

Bajo la hipótesis de simetría completa, la frecuencia esperada de una casilla se obtiene simplemente promediando la frecuencia observada de esa casilla y la frecuencia de su casilla simétrica:

$$\hat{m}_{ij} = (n_{ij} + n_{ji})/2 \quad [8.27]$$

Para poder contrastar la hipótesis de simetría mediante el procedimiento **Loglineal > General** es necesario reorganizar las frecuencias de la tabla de una forma particular. Los datos de la Tabla 8.22 pueden interpretarse como agrupados en dos triángulos. El triángulo inferior contiene los casos en los que el código asignado al estímulo elegido en primer lugar es mayor que el código asignado al estímulo elegido en segundo lugar (ver Tabla 8.25.a). El triángulo superior contiene los casos en los que el código asignado al estímulo elegido en primer lugar es mayor que el código asignado al estímulo elegido en segundo lugar (ver Tabla 8.25.b). Las frecuencias de la diagonal principal no están incluidas en ninguno de los dos triángulos porque no intervienen en la hipótesis de simetría).

Al reordenar las frecuencias de la Tabla 8.22 en dos triángulos, lo que en principio era una tabla bidimensional (con dimensiones *primera elección* y *segunda elección*) se

ha convertido en una tabla tridimensional (con dimensiones *triángulo*, *primera elección* y *segunda elección*). La hipótesis de simetría afirma que las frecuencias de ambos triángulos son iguales. Para contrastar esta hipótesis es necesario introducir los datos de una forma peculiar. La Figura 8.5 muestra el *Editor de datos* con las variables necesarias para reproducir los datos de las Tablas 8.25.a y 8.25.b.

Tabla 8.25.a. Triángulo inferior de la Tabla 8.22

<i>Primera elección</i>	<i>Segunda elección</i>		
	1 = a	2 = b	3 = c
2 = b	14	–	–
3 = c	23	92	–
4 = d	15	38	48

Tabla 8.25.b. Triángulo superior de la Tabla 8.22

<i>Segunda elección</i>	<i>Primera elección</i>		
	1 = a	2 = b	3 = c
2 = b	19	–	–
3 = c	28	89	–
4 = d	14	42	66

Figura 8.5. Datos de las Tablas 8.25.a y 8.25.b reproducidos en el *Editor de datos*

	triángulo	primera	segunda	ncasos	casillas
1	Inferior	b	a	14	1
2	Inferior	b	b	0	0
3	Inferior	b	c	0	0
4	Inferior	c	a	23	1
5	Inferior	c	b	92	1
6	Inferior	c	c	0	0
7	Inferior	d	a	15	1
8	Inferior	d	b	38	1
9	Inferior	d	c	48	1
10	Superior	b	a	19	1
11	Superior	b	b	0	0
12	Superior	b	c	0	0
13	Superior	c	a	28	1
14	Superior	c	b	89	1
15	Superior	c	c	0	0
16	Superior	d	a	14	1
17	Superior	d	b	42	1
18	Superior	d	c	66	1

La variable *triángulo* indica a cuál de los dos triángulos pertenece cada casilla; la variable *primera* recoge los valores de las filas de ambos triángulos; la variable *segunda* recoge los valores de las columnas de ambos triángulos; la variable *ncasos* contiene las frecuencias de las casillas; y la variable *casillas* indica a qué casillas corresponden ceros estructurales (*casillas* = 0). Conviene advertir que los valores de las variables *primera* y *segunda* están codificados de una forma especial: los valores de *primera* se refieren

a la primera elección cuando se trata del triángulo inferior y a la segunda elección cuando se trata del triángulo superior (*primera* recoge las *filas* de ambos triángulos); y los valores de la variable *segunda* se refieren a la segunda elección cuando se trata del triángulo inferior y a la primera elección cuando se trata del superior (*segunda* recoge las *columnas* de ambos triángulos).

Así las cosas, el modelo loglineal de simetría referido a los datos de las Tablas 8.25.a y 8.25.b puede formularse de la siguiente manera:

$$\log_e(m_{ijk}) = \lambda + \lambda_{\text{primera}} + \lambda_{\text{segunda}} + \lambda_{\text{primera} \times \text{segunda}} \quad [8.28]$$

(con $\lambda_{\text{primera}} = \lambda_{\text{segunda}}$). Aunque la variable *triángulo* forma parte de la tabla de contingencias, no está representada por ninguno de los términos del modelo. Esto significa que lo que realmente se está pronosticando con la hipótesis de simetría completa es que el efecto de los términos incluidos en el modelo (*primera*, *segunda* y *primera* $\times$ *segunda*) es el mismo en ambos triángulos, lo cual implica afirmar que las frecuencias de ambos triángulos son iguales. Para ajustar el *modelo loglineal de simetría* a los datos de la Tabla 8.22 (reestructurados en las Tablas 8.25.a y 8.25.b):

- Reproducir los datos de la Tabla 8.22 tal como muestra la Figura 8.5 y ponderar el archivo con la variable *ncasos* mediante la opción **Ponderar casos** del menú **Datos** (o descargar el archivo *Loglineal simetría completa* de la página web del manual).
- En el cuadro de diálogo *Análisis loglineal general*, trasladar las variables *triángulo*, *primera* y *segunda* a la lista **Factores** y la variable *casillas* al cuadro **Estructura de las casillas**.
- Pulsar el botón **Modelo** para acceder al subcuadro de diálogo *Análisis loglineal general: Modelo*, marcar la opción **Personalizado** e incorporar a la lista **Términos del modelo** los dos efectos principales *primera* y *segunda* y la interacción *primera* $\times$ *segunda*. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones se obtienen, entre otros, los resultados que muestran las Tablas 8.26 y 8.27. En la primera de ellas se puede constatar que las casillas con ceros estructurales están correctamente estimadas: todas las frecuencias esperadas de esas casillas valen cero. Y puede constatare también que las frecuencias esperadas de las casillas no vacías son idénticas en ambos triángulos. Por ejemplo, la frecuencia esperada de la casilla 2-1 del triángulo inferior vale 16,5, lo mismo que la casilla 2-1 del triángulo superior; la frecuencia esperada de la casilla 3-1 del triángulo inferior vale 25,5, lo mismo que la casilla 3-1 del triángulo superior; etc. Es decir, las frecuencias esperadas reflejan las predicciones del modelo de simetría, que es justamente el que estamos intentando ajustar.

La Tabla 8.27 ofrece los dos estadísticos de bondad de ajuste. El nivel crítico asociado a la razón de verosimilitudes ($\text{sig.} = 0,624 > 0,05$) indica que el modelo de simetría ofrece un buen ajuste a los datos y, consecuentemente, que las frecuencias de ambos triángulos siguen una pauta similar. No parece, por tanto, que la frecuencia con la que es elegido el estímulo de cada par en primer lugar sea distinta de la frecuencia con la que es elegido en segundo lugar.

Los residuos tipificados corregidos y los de desviación también indican que el modelo de simetría completa ofrece un buen ajuste a los datos: todos ellos toman valores comprendidos entre $-1,96$ y $1,96$.

Tabla 8.26. Frecuencias y residuos (simetría completa)

Triáng.	Estímulo primera elección	Estímulo segunda elección	Observado		Esperado		Resid.	Resid. tipificad.	Resid. corregid.	Resid. desvian.
			n	%	n	%				
Inferior	b	a	14	2,9%	16,50	3,4%	-2,50	-,62	-,87	-,63
		b	0	,0%	,00	,0%	.	.	.	.
		c	0	,0%	,00	,0%	.	.	.	.
	c	a	23	4,7%	25,50	5,2%	-2,50	-,50	-,70	-,50
		b	92	18,9%	90,50	18,5%	1,50	,16	,22	,16
		c	0	,0%	,00	,0%	.	.	.	.
	d	a	15	3,1%	14,50	3,0%	,50	,13	,19	,13
		b	38	7,8%	40,00	8,2%	-2,00	-,32	-,45	-,32
		c	48	9,8%	57,00	11,7%	-9,00	-1,19	-1,69	-1,23
Superior	b	a	19	3,9%	16,50	3,4%	2,50	,62	,87	,60
		b	0	,0%	,00	,0%	.	.	.	.
		c	0	,0%	,00	,0%	.	.	.	.
	c	a	28	5,7%	25,50	5,2%	2,50	,50	,70	,49
		b	89	18,2%	90,50	18,5%	-1,50	-,16	-,22	-,16
		c	0	,0%	,00	,0%	.	.	.	.
	d	a	14	2,9%	14,50	3,0%	-,50	-,13	-,19	-,13
		b	42	8,6%	40,00	8,2%	2,00	,32	,45	,31
		c	66	13,5%	57,00	11,7%	9,00	1,19	1,69	1,16

Tabla 8.27. Estadísticos de bondad de ajuste (simetría completa)

	Valor	gl	Sig.
Razón de verosimilitudes	4,39	6	,624
Chi-cuadrado de Pearson	4,37	6	,626

Simetría relativa

El modelo de simetría completa permite contrastar la hipótesis nula de que las frecuencias de ambos triángulos son iguales, es decir, la hipótesis de que la probabilidad de la casilla ij del triángulo inferior es idéntica a la probabilidad de la casilla ji del triángulo superior. Por tanto, en la simetría completa no se contempla la posibilidad de que las frecuencias totales de uno de los dos triángulos puedan ser mayores que las del otro, es decir, no se contempla la posibilidad de que pertenecer a uno de los dos triángulos pueda ser más probable que pertenecer al otro. Cuando se desea tener en cuenta esta circunstancia, el modelo de simetría completa no sirve para contrastar la hipótesis de que las probabilidades de los dos triángulos son simétricamente iguales.

Las Tablas 8.28.a y 8.28.b ilustran una situación en la que el tamaño de las frecuencias del triángulo superior son sensiblemente mayores que las del triángulo inferior. La primera de ellas reproduce las frecuencias de la Tabla 8.28.a (ambas son idénticas) y la segunda reproduce las frecuencias de la Tabla 8.28.b multiplicadas por 3 para forzar que las frecuencias del segundo triángulo sean mayores que las del primero.

Tabla 8.28.a. Frecuencias del triángulo inferior

<i>Primera elección</i>	<i>Segunda elección</i>		
	1 = a	2 = b	3 = c
2 = b	14	–	–
3 = c	23	92	–
4 = d	15	38	48

Tabla 8.28.b. Frecuencias del triángulo superior

<i>Segunda elección</i>	<i>Primera elección</i>		
	1 = a	2 = b	3 = c
2 = b	57	–	–
3 = c	84	267	–
4 = d	42	126	198

Al ajustar el modelo de simetría completa a estos datos se obtiene, para la razón de verosimilitudes, un nivel crítico menor que 0,0005, lo que indica que el modelo de simetría completa no ofrece un buen ajuste a los datos (lo cual no es sorprendente dada la enorme diferencia existente entre las frecuencias de ambos triángulos).

Cuando en una situación de estas características todavía sigue interesando valorar si las frecuencias de ambos triángulos siguen o no la misma pauta, lo apropiado es ajustar el modelo de **simetría relativa**. Este modelo sigue hipotetizando que la probabilidad de la casilla ij es la misma en ambos triángulos, pero tiene en cuenta que el tamaño de las frecuencias de los triángulos puede ser distinto. El modelo loglineal de simetría relativa adopta la siguiente forma:

$$\log_e(m_{ijk}) = \lambda + \lambda_{\text{triángulo}} + \lambda_{\text{primera}} + \lambda_{\text{segunda}} + \lambda_{\text{primera} \times \text{segunda}} \quad [8.29]$$

Este modelo es idéntico al de simetría completa propuesto en [8.28], excepto por lo que se refiere al término que recoge el efecto de la variable *triángulo*, cuya presencia en el modelo permite tener en cuenta el tamaño de los triángulos.

Ahora, las frecuencias esperadas no se estiman a partir de la media entre la frecuencia de una casilla y la frecuencia de la casilla simétricamente opuesta (que es como se hace para ajustar la hipótesis de simetría completa), sino con la *media ponderada*: la frecuencia esperada de una casilla se estima a partir del producto entre la probabilidad de esa casilla y el tamaño total del triángulo al que pertenece. Para ajustar el modelo de simetría relativa a los datos de las Tablas 8.28.a y 8.28.b:

- Reproducir los datos de las Tablas 8.28.a y 8.28.b tal como se ha hecho en el apartado anterior con los datos de las Tablas 8.25.a y 8.25.b y ponderar el archivo con la variable *ncasos* mediante la opción **Ponderar casos** del menú **Datos** (o descargar el archivo *Loglineal simetría relativa* que se encuentra en la página web del manual).

- En el cuadro de diálogo *Análisis loglineal general*, trasladar las variables *triángulo*, *primera* y *segunda* a la lista **Factores** y la variable *casillas* al cuadro **Estructura de las casillas**.
- Pulsar el botón **Modelo** para acceder al subcuadro de diálogo *Análisis loglineal general: Modelo*, marcar la opción **Personalizado** e incorporar a la lista **Términos del modelo** los tres efectos principales *triángulo*, *primera* y *segunda* y la interacción *primera* × *segunda*. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Con estas selecciones se obtienen, entre otros, los resultados de las Tablas 8.29 y 8.30. En la primera de ellas puede constatar, de nuevo, que las casillas con ceros estructurales están correctamente estimadas (las frecuencias esperadas de todas esas casillas valen cero). Pero existe una diferencia sustancial entre estos resultados y los obtenidos al contrastar la hipótesis de simetría completa (ver Tabla 8.26): ahora, las frecuencias esperadas de las casillas no vacías no son idénticas en ambos triángulos. Por ejemplo, mientras que la frecuencia esperada de la casilla 2-1 del triángulo inferior vale 16,26, la frecuencia esperada de la casilla 2-1 del triángulo superior vale 54,74.

La Tabla 8.30 ofrece los dos estadísticos de ajuste. Puesto que el nivel crítico asociado a la razón de verosimilitudes ($\text{sig.} = 0,54$) es mayor que 0,05, puede concluirse que el modelo de simetría relativa ofrece un buen ajuste a los datos. Por tanto, cuando se controla el tamaño de los triángulos, parece que las frecuencias de ambos triángulos muestran una pauta de variación similar.

Tabla 8.29. Frecuencias y residuos (simetría relativa)

Triáng.	Estímulo primera elección	Estímulo segunda elección	Observado		Esperado		Residuo s	Resid. tipificad.	Resid. corregid.	Resid. desvian.
			n	%	n	%				
Inferior	b	a	14	1,4%	16,26	1,6%	-2,26	-,56	-,66	-,58
		b	0	,0%	,00	,0%	.	.	.	.
		c	0	,0%	,00	,0%	.	.	.	.
	c	a	23	2,3%	24,51	2,4%	-1,51	-,31	-,37	-,31
		b	92	9,2%	82,24	8,2%	9,76	1,08	1,53	1,06
		c	0	,0%	,00	,0%	.	.	.	.
	d	a	15	1,5%	13,06	1,3%	1,94	,54	,63	,52
		b	38	3,8%	37,57	3,7%	,43	,07	,09	,07
		c	48	4,8%	56,35	5,6%	-8,35	-1,11	-1,46	-1,14
Superior	b	a	57	5,7%	54,74	5,5%	2,26	,31	,66	,30
		b	0	,0%	,00	,0%	.	.	.	.
		c	0	,0%	,00	,0%	.	.	.	.
	c	a	84	8,4%	82,49	8,2%	1,51	,17	,37	,17
		b	267	26,6%	276,76	27,6%	-9,76	-,59	-1,53	-,59
		c	0	,0%	,00	,0%	.	.	.	.
	d	a	42	4,2%	43,94	4,4%	-1,94	-,29	-,63	-,30
		b	126	12,5%	126,43	12,6%	-,43	-,04	-,09	-,04
		c	198	19,7%	189,65	18,9%	8,35	,61	1,46	,60

Tabla 8.30. Estadísticos de bondad de ajuste (simetría relativa)

	Valor	gl	Sig.
Razón de verosimilitudes	4,05	5	,543
Chi-cuadrado de Pearson	4,02	5	,547

Cuando los modelos de simetría completa y simetría relativa no ofrecen un buen ajuste a los datos, todavía es posible formular otros modelos que imponen sobre los datos menos restricciones. Uno de estos modelos es el de **cuasi-simetría corregida**, el cual permite estudiar si la pauta que siguen las frecuencias de ambos triángulos es la misma cuando, además de tener en cuenta que el tamaño de los triángulos puede ser distinto (hipótesis de simetría relativa), se considera que las frecuencias marginales de los triángulos también pueden ser distintas.

La hipótesis de *simetría completa* pronostica que las frecuencias de las casillas simétricamente opuestas son iguales. La hipótesis de *simetría relativa* pronostica que las frecuencias de las casillas simétricamente opuestas son proporcionalmente iguales, es decir, iguales cuando se tiene en cuenta que el tamaño de los dos triángulos puede ser distinto. La hipótesis de *cuasi-simetría corregida* pronostica que las frecuencias de ambos triángulos son condicionalmente iguales, es decir, iguales cuando se tiene en cuenta que las frecuencias marginales de los dos triángulos pueden ser distintas. El modelo loglineal de cuasi-simetría corregida adopta la siguiente forma:

$$\log_e(m_{ijk}) = \lambda + \lambda_{\text{triángulo}} + \lambda_{\text{primera}} + \lambda_{\text{segunda}} + \lambda_{\text{triángulo} \times \text{primera}} + \lambda_{\text{triángulo} \times \text{segunda}} + \lambda_{\text{primera} \times \text{segunda}} \quad [8.30]$$

La diferencia entre este modelo y el saturado está en que el modelo de cuasi-simetría corregida no incluye el término referido a la interacción triple *triángulo* $\times$ *primera* $\times$ *segunda*. Para ajustar el modelo de cuasi-simetría corregida a los datos de las Tablas 8.28.a y 8.28.b basta con repetir los pasos seguidos para ajustar el modelo de simetría relativa del último ejemplo, pero cambiando un detalle: en la lista **Términos del modelo**, además de los efectos *triángulo*, *primera*, *segunda* y *primera* $\times$ *segunda*, hay que incluir las interacciones *triángulo* $\times$ *primera* y *triángulo* $\times$ *segunda*.

Tasas de respuesta

Aunque todos los modelos loglineales estudiados hasta ahora sirven para pronosticar *frecuencias absolutas* (número de casos), los modelos loglineales también pueden utilizarse para analizar tablas de contingencias cuando el contenido de las casillas no son frecuencias absolutas sino *tasas de respuesta*.

Una *tasa* es un número de eventos de algún tipo dividido por una línea base relevante. Por ejemplo, el número de fumadores de más de 20 cigarrillos/día que padecen cáncer de pulmón dividido por el inverso del tiempo de exposición al tabaco; o el número de accidentes de tráfico dividido por la cantidad de vehículos que circulan durante

un año, o el número de muertes que se producen al año dividido por el número de habitantes, etc. Cuando se trabaja con *tasas*, las casillas de la tabla contienen dos valores: el *número de eventos* (n_{ij}) y el *denominador de la tasa* (N_{ij} : tiempo de exposición, número de vehículos, horas de funcionamiento, etc.).

Recordemos que el modelo loglineal saturado referido a las frecuencias de una tabla de contingencias bidimensional adopta la forma (ver ecuación [8.7]):

$$\log_e(m_{ij}) = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{XY(ij)}$$

Ahora bien, si en lugar de *frecuencias absolutas* (m_{ij}) interesa estudiar *tasas de res puesta* (es decir, m_{ij}/N_{ij}), el modelo loglineal saturado adopta la siguiente forma:

$$\log_e(m_{ij}/N_{ij}) = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{XY(ij)} \quad [8.31]$$

es decir:

$$\log_e(m_{ij}) - \log_e(N_{ij}) = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{XY(ij)} \quad [8.32]$$

Al término $\log_e(N_{ij})$ se le suele llamar *término de compensación* (*offset*; ver, por ejemplo, Agresti, 1990, pág. 193).

La Tabla 8.31 muestra los datos de una aseguradora referidos al número de accidentes que han sufrido sus vehículos asegurados durante un año. La tabla incluye la *antigüedad del vehículo*, la *edad del conductor*, la *cilindrada del vehículo*, el *número de accidentes observados* y el *número de vehículos expuestos*. Las variables *antigüedad*, *cilindrada* y *edad* reproducen los patrones de variabilidad. Las variables *accidentes* y *expuestos* recogen el número de accidentes y el número de vehículos expuestos en cada patrón de variabilidad.

Tabla 8.31. Tabla de contingencias de *antigüedad* por *cilindrada* por *edad del conductor*

<i>Antigüedad del vehículo</i>	<i>Cilindrada del vehículo</i>	<i>Edad del conductor</i>			
		Hasta 25 años		Más de 25 años	
		nº accid.	nº vehíc.	nº accid.	nº vehíc.
Hasta 10 años	Hasta 2.000cc	15422	902271	16707	145711
	Más de 2.000cc	12115	359513	13702	608662
Más de 10 años	Hasta 2.000cc	8584	157502	3633	98023
	Más de 2.000cc	3312	30781	14549	197038

A los datos de la Tabla 8.31 puede ajustarse cualquier modelo loglineal de los ya estudiados. Supongamos que se desea ajustar el modelo de independencia. Es decir, supongamos que los vehículos más antiguos tienen más accidentes independientemente de su cilindrada y de la edad del conductor; que los vehículos de mayor cilindrada tienen más

accidentes independientemente de su antigüedad y de la edad del conductor; y que los conductores más jóvenes tienen más accidentes independientemente de la antigüedad y de la cilindrada del vehículo. Para ajustar este modelo loglineal de independencia a los datos de la Tabla 8.31:

- Reproducir los datos de la Tabla 8.31 tal como muestra la Figura 8.6 y ponderar el archivo con la variable *accidentes* mediante la opción **Ponderar casos** del menú **Datos** (o descargar el archivo *Loglineal tasas de respuesta* que se encuentra en la página web del manual).

Figura 8.6. Datos de la Tabla 8.31 reproducidos en el *Editor de datos*

	antigüedad	cilindrada	edad	accidentes	expuestos
1	Hasta 10 años	Hasta 2.000 cc	Hasta 25 años	15422	902271
2	Hasta 10 años	Hasta 2.000 cc	Más de 25 años	16707	1457112
3	Hasta 10 años	Más de 2.000 cc	Hasta 25 años	12115	359513
4	Hasta 10 años	Más de 2.000 cc	Más de 25 años	13702	608662
5	Más de 10 años	Hasta 2.000 cc	Hasta 25 años	8584	157502
6	Más de 10 años	Hasta 2.000 cc	Más de 25 años	3633	98023
7	Más de 10 años	Más de 2.000 cc	Hasta 25 años	3312	30781
8	Más de 10 años	Más de 2.000 cc	Más de 25 años	14549	197038

- En el cuadro de diálogo *Análisis loglineal general*, trasladar las variables *antigüedad*, *cilindrada* y *edad* a la lista **Factores** y la variable *expuestos* al cuadro **Estructura de las casillas**.
- Pulsar el botón **Modelo** para acceder al subcuadro de diálogo *Análisis loglineal general: Modelo*, marcar la opción **Personalizado** e incorporar a la lista **Términos del modelo** los tres efectos principales *antigüedad*, *cilindrada* y *edad*. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Opciones** para acceder al subcuadro de diálogo *Análisis loglineal general: Opciones* y marcar la opción **Estimaciones** del recuadro **Mostrar**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, se obtienen, entre otros, los resultados que muestran las Tablas 8.32 y 8.33. Puesto que el nivel crítico asociado a la razón de verosimilitudes vale 0,57 (ver Tabla 8.33), se puede concluir que el modelo de independencia ofrece un buen ajuste a los datos. El tamaño de los residuos corregidos apunta en la misma dirección: todos ellos toman valores comprendidos entre -1,96 y 1,96.

Aunque el número de vehículos expuestos no aparece entre los resultados, el ajuste se ha realizado teniendo en cuenta la estructura de las casillas. Las estimaciones de los parámetros permiten comprobar esta circunstancia. El modelo loglineal de independencia propuesto para estudiar la tasa de accidentes adopta la forma:

$$\log_e(m_{ijk}) - \log_e(N_{ijk}) = \lambda + \lambda_{\text{antigüedad}(i)} + \lambda_{\text{cilindrada}(j)} + \lambda_{\text{edad}(k)}$$

Tabla 8.32. Frecuencias y residuos

Antigüedad del vehículo	Cilindrada del vehículo	Edad del conductor	Observado		Esperado		Resid.	Resid. corregid.
			n	%	n	%		
Hasta 10 años	Hasta 2.000 cc	Hasta 25 años	15422	17,5%	15343,24	17,4%	78,76	,95
		Más de 25 años	16707	19,0%	16712,12	19,0%	-5,12	-,07
	Más de 2.000 cc	Hasta 25 años	12115	13,8%	12087,81	13,7%	27,19	,39
		Más de 25 años	13702	15,6%	13802,82	15,7%	-100,82	-1,25
Más de 10 años	Hasta 2.000 cc	Hasta 25 años	8584	9,8%	8656,84	9,8%	-72,84	-1,14
		Más de 25 años	3633	4,1%	3633,79	4,1%	-,79	-,01
	Más de 2.000 cc	Hasta 25 años	3312	3,8%	3345,10	3,8%	-33,10	-,64
		Más de 25 años	14549	16,5%	14442,26	16,4%	106,74	1,53

Tabla 8.33. Estadísticos de bondad de ajuste

	Valor	gl	Sig.
Razón de verosimilitudes	2,93	4	,569
Chi-cuadrado de Pearson	2,93	4	,569

La frecuencias esperadas pueden obtenerse a partir de las estimaciones de los parámetros que ofrece la Tabla 8.34. Por ejemplo, la primera casilla de la tabla (definida por la combinación *antigüedad* = “hasta 10 años”, *cilindrada* = “hasta 2.000 cc”, *edad* = “hasta 25 años”), se obtiene mediante:

$$\log_e(\hat{m}_{111}) = \log_e(N_{111}) + \hat{\lambda} + \hat{\lambda}_{\text{antigüedad}(1)} + \hat{\lambda}_{\text{cilindrada}(1)} + \hat{\lambda}_{\text{edad}(1)}$$

Tomando el valor de N_{111} y sustituyendo cada parámetro lambda por su correspondiente estimación se obtiene el logaritmo de la frecuencia esperada para la casilla 1-1-1:

$$\log_e(\hat{m}_{111}) = \log_e(902.271) - 2,6132 - 1,1732 - 0,6817 + 0,3938 = 9,6384$$

Por tanto: $\hat{m}_{111} = e^{9,6384} = 15.342,78$ (valor que, salvo por pequeñas diferencias debidas al redondeo, coincide con la frecuencia estimada por el modelo para esa casilla).

Tabla 8.34. Estimaciones de los parámetros

Parámetro	Estimación	Error típico	Z	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Constante	-2,6132	,01	-385,27	,000	-2,63	-2,60
[antigüedad = 1]	-1,1732	,01	-163,36	,000	-1,19	-1,16
[antigüedad = 2]	,0000 ^a	.	.	.	.	.
[cilindrada = 1]	-,6817	,01	-98,51	,000	-,70	-,67
[cilindrada = 2]	,0000 ^a	.	.	.	.	.
[edad = 1]	,3938	,01	57,00	,000	,38	,41
[edad = 2]	,0000 ^a	.	.	.	.	.

a. Este parámetro se ha definido como cero ya que es redundante.

Los valores exponenciales de los coeficientes pueden interpretarse como si fueran *odds ratios* (aunque referidas a las *tasas* de respuesta, no a las *odds*). Por ejemplo, el valor estimado para *antigüedad* = 1 es el logaritmo del cociente entre la tasa de accidentes para esa antigüedad y la tasa de accidentes para *antigüedad* = 2 (que es la categoría de referencia de esa variable, es decir, la categoría cuyo parámetro se ha fijado en cero). Por tanto, $e^{-1,1732} = 0,31$ indica que la tasa de accidentes pronosticada para los vehículos con menos de 10 años es un 31% de la tasa de accidentes pronosticada para los vehículos con más de 10 años (lo cual equivale a decir que la primera tasa es un 69% menor que la segunda). Puesto que la antigüedad del vehículo es independiente del resto de variables, esta afirmación es válida para todas las cilindradas y edades.

Por el mismo razonamiento, el valor estimado para *edad* = 1 es el logaritmo del cociente entre la tasa de accidentes para esa edad y la tasa de accidentes para *edad* = 2. Por tanto, $e^{0,3928} = 1,48$ indica que la tasa de accidentes pronosticada para los conductores de 25 años o menos es un 48% más alta que la tasa de accidentes pronosticada para los conductores de más de 25 años. Y dado que la edad no parece estar relacionada con el resto de variables, esta afirmación es válida para todas las antigüedades y cilindradas.

Comparaciones entre niveles

Cuando una variable categórica posee un efecto significativo existe la posibilidad de comparar sus niveles o categorías para averiguar cuáles difieren entre sí. Para poder hacer esto es necesario crear una variable con códigos (unos y ceros) que definan la comparación que se desea realizar. Estos códigos se asignan con la misma lógica con la que se asignan los códigos a los niveles de un factor para definir comparaciones planeadas en un análisis de varianza.

Retomemos los datos de la Tabla 8.22, reproducidos en la Figura 8.7, y supongamos que estamos interesados en averiguar cuál es el estímulo más elegido en primera opción. Aunque las frecuencias marginales de la tabla indican que los estímulos más elegidos son el *b* y el *c*, podría interesar comparar esos estímulos entre sí para averiguar si la diferencia de elección observada es o no significativa. Para ello, puesto que la tabla tiene 16 casillas, es necesario crear una variable con 16 códigos; por ejemplo:

0, 0, 0, 0, 1, 1, 1, 1, -1, -1, -1, -1, 0, 0, 0, 0, 0, 0, 0, 0

Los cuatro primeros códigos corresponden a las cuatro casillas en las que se ha elegido en primera opción el estímulo *a*; los cuatro siguientes, a las cuatro casillas en las que se ha elegido en primera opción el estímulo *b*; ..., los cuatro últimos, a las cuatro casillas en las que se ha elegido en primera opción el estímulo *d*. La comparación se efectúa entre las casillas con código positivo y las casillas con código negativo. Las casillas con código igual a cero no intervienen en la comparación.

Por supuesto, estos códigos pueden asignarse de tal forma que, en lugar de definir una comparación entre los niveles de una variable *factor*, se defina una comparación entre casillas concretas. Por ejemplo, para averiguar, cuando se eligen los estímulos *a*

y *b*, cuál de ellos es más elegido en primera opción, podrían asignarse ceros a todas las casillas excepto a la segunda (1) y a la quinta (-1).

La Figura 8.7 muestra las variables del archivo original (Figura 8.4) más dos variables nuevas: los códigos de la primera de ellas (*comp_1*) definen la comparación entre los estímulos *b* y *c* en la primera elección; los códigos de la segunda variable (*comp_2*) definen la comparación entre los estímulos *b* y *c* en la segunda elección (estos datos están disponibles en el archivo *Loglineal comparar niveles* en la página web del manual).

Figura 8.7. Datos de la Tabla 8.22 (ver Figura 8.4) más dos variables de contraste

	primera	segunda	ncasos	casillas	comp_1	comp_2
1	a	a	0	0	0	0
2	a	b	19	1	0	1
3	a	c	28	1	0	-1
4	a	d	14	1	0	0
5	b	a	14	1	1	0
6	b	b	0	0	1	1
7	b	c	89	1	1	-1
8	b	d	42	1	1	0
9	c	a	23	1	-1	0
10	c	b	92	1	-1	1
11	c	c	0	0	-1	-1
12	c	d	66	1	-1	0
13	d	a	15	1	0	0
14	d	b	38	1	0	1
15	d	c	48	1	0	-1
16	d	d	0	0	0	0

A estos datos ya hemos ajustado el modelo loglineal de *cuasi-independencia* (ver el apartado *Tablas cuadradas: Cuasi-independencia*). Veamos ahora cómo comparar los estímulos *b* y *c* en la primera elección y en la segunda:

- En el cuadro de diálogo *Análisis loglineal general*, trasladar las variables *primera* y *segunda* a la lista **Factores**, la variable *casillas* al cuadro **Estructura de las casillas** y las variables *comp_1* y *comp_2* a la lista **Variables de contraste**.
- Pulsar el botón **Modelo** para acceder al subcuadro de diálogo *Análisis loglineal general: Modelo*, marcar la opción **Personalizado** y definir, como **Términos del modelo**, los dos efectos principales *primera* y *segunda*.

Al incluir una o más variables de contraste, el *Visor* ofrece, además de los resultados ya conocidos (ajuste del modelo de cuasi-independencia, frecuencias esperadas y residuos, etc.), la *odds ratio generalizada* (el SPSS la llama *razón de las ventajas generalizada*; ver Tabla 8.35). Para la primera comparación (*comp_1*) se obtiene un valor estimado de -0,874. Este valor refleja, en escala logarítmica, el cociente entre la *odds*

estimada para el estímulo *b* y la *odds* estimada para el estímulo *c* (ambos referidos a la primera elección). Este valor indica (ver frecuencias esperadas de la Tabla 10.15; debe tenerse en cuenta que las casillas 2-2 y 3-3 no intervienen en el análisis) que

$$\log_e \left(\frac{\hat{m}_{21} \hat{m}_{23} \hat{m}_{24}}{\hat{m}_{31} \hat{m}_{32} \hat{m}_{34}} \right) = \log_e \left(\frac{16,70 \times 84,96 \times 43,34}{24,94 \times 91,35 \times 64,71} \right) = -0,87$$

El valor exponencial del coeficiente, $e^{-0,87} = 0,42$, es el valor que el modelo de cuasi-independencia estima como indicador del grado en que el estímulo *b* es más elegido que el estímulo *c*. Por tanto, el estímulo *b* es elegido, en primera opción, un 42% de lo que es elegido el estímulo *c*. O, de otra forma, que el estímulo *c* es elegido en primera opción $1/0,42 = 2,38$ veces más que el *b*. Y tanto el nivel crítico (*sig.* = 0,012 > 0,05) como los correspondientes intervalos de confianza indican que esta diferencia es significativamente distinta de cero.

Por lo que se refiere a la segunda comparación (*comp_2*), el valor estimado para el logaritmo de la *odds ratio generalizada* es -0,58. Su valor exponencial, $e^{-0,58} = 0,56$, indica que (siempre según las estimaciones del modelo de cuasi-independencia) el estímulo *b* es elegido, en segunda opción, un 56% de lo que es elegido el estímulo *c*. O, de otra forma, que el estímulo *c* es elegido en segunda opción $1/0,56 = 1,79$ veces más que el *b*. Pero tanto el nivel crítico (*sig.* = 0,098) como los correspondientes intervalos de confianza indican que esta diferencia no alcanza la significación estadística.

Tabla 8.35. Razones de las ventajas (*odds ratios*) generalizadas

		Valor	Error típico	Wald	Sig.	Intervalo de confianza al 95%	
						L. inferior	L. superior
comp_1	Log-razón de las ventajas generalizadas	-,87	,35	6,36	,012	-1,55	-,19
	Razón de las ventajas generalizadas	,42				,21	,82
comp_2	Log-razón de las ventajas generalizadas	-,58	,35	2,74	,098	-1,28	,11
	Razón de las ventajas generalizadas	,56				,28	1,11

Modelos logit

Los **modelos loglineales** estudiados hasta ahora no distinguen entre variables dependientes e independientes. Todas ellas se consideran independientes; la variable dependiente de un modelo loglineal es el contenido de las casillas (número de veces que se repite cada patrón de variabilidad); y el objetivo del análisis es explorar la pauta de asociación existente entre las variables. En los **modelos logit** se distingue entre variables dependientes e independientes; y el objetivo del análisis es describir el efecto de una o más variables independientes sobre una variable dependiente (todas ellas categóricas).

Los modelos logit no son una cosa distinta de la **regresión logística** ya estudiada en el Capítulo 5. No obstante, la regresión logística trabaja con los datos no agrupados

(admite variables cuantitativas), mientras que los modelos logit utilizan una estrategia basada en la agrupación de casos. Esta diferencia en la forma de tratar los datos genera modelos saturados distintos. Y de ahí derivan las diferencias entre ambos enfoques.

Los modelos logit pueden considerarse una versión particular de los loglineales, pero la diferencia entre ambos es importante. En los modelos loglineales se modelan las *frecuencias* de una tabla de contingencias; en los logit se modela el *cociente entre las frecuencias de dos categorías* de la variable que se toma como dependiente. Por ejemplo, al estudiar la relación entre las variables *tratamiento* (A, B) y *recuperación* (sí, no), un modelo loglineal modela las frecuencias de las cuatro casillas resultantes de cruzar ambas variables; si la variable *recuperación* se toma como dependiente, un modelo logit modela el cociente (la *odds*) entre las frecuencias de las dos categorías de la variable recuperación¹⁰.

Una variable independiente

Los pasos que es necesario seguir para formular y evaluar un modelo logit son los ya descritos a propósito de los modelos loglineales. Por esta razón, algunas de las cuestiones ya tratadas serán pasadas por alto aquí. Comencemos considerando una situación con dos variables (X e Y) donde la variable X es una variable categórica (con categorías $i = 1, 2, \dots, I$) que se toma como *independiente* o *predictora* y la variable Y es una variable dicotómica (con categorías $j = 1, 2$) que se toma como *dependiente* o *respuesta*. Puesto que Y es una variable dicotómica, la probabilidad condicional de que Y tome el valor $j = 1$ en cada categoría de X viene dada por

$$P(Y = 1 | X = i) = \pi_{i1} / (\pi_{i1} + \pi_{i2}) = \pi_{i1} / \pi_{i+} \quad [8.33]$$

Tomando frecuencias absolutas en lugar de relativas y desarrollando, el *logit* de Y en cada categoría de X puede formularse como

$$\text{logit}(Y = 1) = \log_e(m_{i1} / m_{i2}) \quad [8.34]$$

Recordemos que, con dos variables, el **modelo loglineal saturado**, es decir, el modelo que incluye todos los términos posibles, adopta la siguiente forma (ver ecuación [8.7]):

$$\log_e(m_{ij}) = \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{XY(ij)}$$

Pero con un modelo logit no se está modelando el contenido de las casillas, m_{ij} , sino el cociente entre las frecuencias correspondientes a las dos categorías de la variable dependiente. En consecuencia¹¹,

$$\text{logit}(Y = 1) = [\lambda_{Y(1)} - \lambda_{Y(2)}] + [\lambda_{XY(i1)} - \lambda_{XY(i2)}] \quad [8.35]$$

¹⁰ Aunque es posible ajustar modelos logit con variables dependientes politómicas, aquí nos limitaremos a estudiar el caso más común: una variable dependiente dicotómica.

¹¹ En efecto, $\text{logit}(Y = 1) = \log_e(m_{i1}) - \log_e(m_{i2}) = [\lambda + \lambda_{X(i)} + \lambda_{Y(1)} + \lambda_{XY(i1)}] - [\lambda + \lambda_{X(i)} + \lambda_{Y(2)} + \lambda_{XY(i2)}]$. Y pues-
to que los términos λ y $\lambda_{X(i)}$ desaparecen (pues están repetidos con signo cambiado), el modelo logit que equivale al modelo loglineal saturado se reduce a: $\text{logit}(Y) = [\lambda_{Y(1)} - \lambda_{Y(2)}] + [\lambda_{XY(i1)} - \lambda_{XY(i2)}]$.

Aplicando ahora una notación similar a la de los modelos loglineales, el **modelo logit** que incluye todos los términos posibles en el contexto de una tabla de contingencias bidimensional puede expresarse como:

$$\text{logit}(Y=1) = \xi + \xi_{X(i)} \quad [8.36]$$

donde $\xi = \lambda_{Y(1)} - \lambda_{Y(2)}$ representa la diferencia entre las dos categorías de la variable Y y $\xi_{X(i)} = \lambda_{XY(i1)} - \lambda_{XY(i2)}$ representa la asociación entre la variable independiente y la variable dependiente. Es importante reparar en el hecho de que un modelo logit no incluye ningún término referido a la variable independiente individualmente considerada; únicamente incluye los términos en los que está involucrada la variable dependiente.

Los términos λ_h suman cero para cada efecto (pues se conciben como desviaciones del promedio o de alguna categoría de referencia). Como, además, la variable Y es dicotómica, se verifica

$$\sum_j \lambda_{Y(j)} = \sum_i \lambda_{XY(i1)} = \sum_i \lambda_{XY(i2)} = 0 \quad [8.37]$$

Por tanto, el término $\lambda_{Y(j)}$ contiene un solo parámetro independiente y el término $\lambda_{XY(i1)}$ contiene $I - 1$ parámetros independientes. Si estos parámetros se estiman fijando en cero los redundantes (como hacen los procedimientos **General** y **Logit**), se tiene:

$$\begin{aligned} \xi &= \lambda_{Y(1)} - \lambda_{Y(2)} = \lambda_{Y(1)} \\ \xi_{X(i)} &= \lambda_{XY(i1)} - \lambda_{XY(i2)} = \lambda_{XY(i1)} \end{aligned} \quad [8.38]$$

Veamos qué significan estos coeficientes. La Tabla 8.36 recoge el resultado de clasificar una muestra de 240 sujetos en las variables $X = \text{"sexo"}$ e $Y = \text{"consumo de marihuana en el último año"}$. Utilizando el procedimiento **Loglineal > Logit** para estimar los parámetros del modelo saturado (más adelante veremos cómo hacer esto), se obtiene

$$\text{logit}(Y=1) = \xi + \xi_{\text{hombres}} = \hat{\lambda}_{\text{sí}} + \hat{\lambda}_{\text{hombres, sí}} = -0,94 + 1,69 \quad [8.39]$$

El resto de parámetros ($\hat{\lambda}_{\text{hombres, no}}$, $\hat{\lambda}_{\text{mujeres, sí}}$, $\hat{\lambda}_{\text{mujeres, no}}$) son redundantes y su valor se fija en cero. Aplicando estas estimaciones a los datos de la Tabla 8.36 se obtienen dos pronósticos distintos:

$$\begin{aligned} \text{logit}(Y=1 \mid \text{hombres}) &= \hat{\lambda}_{\text{sí}} + \hat{\lambda}_{\text{hombres, sí}} = -0,94 + 1,69 = 0,75 \\ \text{logit}(Y=1 \mid \text{mujeres}) &= \hat{\lambda}_{\text{sí}} + \hat{\lambda}_{\text{mujeres, sí}} = -0,94 + 0 = -0,94 \end{aligned}$$

Tabla 8.36. Frecuencias conjuntas de sexo por consumo de marihuana

(X) Sexo	(Y) Consumo marihuana		
	Sí	No	
Hombres	68	32	$odds(\text{sí} \mid \text{hombres}) = 68/32 = 2,12$
Mujeres	40	102	$odds(\text{sí} \mid \text{mujeres}) = 40/102 = 0,39$

Estos dos pronósticos están en escala logit. Sus valores exponenciales son las *odds* del suceso *consumir marihuana* en el grupo de *hombres* ($e^{0.75} = 2,12$) y en el de *mujeres* ($e^{-0.94} = 0,39$). Y el cociente de esas dos *odds* (*odds ratio* = $2,12/0,39 = 5,44$) coincide, salvo por detalles de redondeo, con $e^{1.69} = 5,42$. Por tanto, los dos coeficientes no redundantes del modelo logit propuesto en [8.39] tienen una interpretación clara:

- Coeficiente $\hat{\lambda}_{st}$. El valor exponencial del término constante es la *odds* del suceso estudiado (consumir marihuana) cuando todas las variables independientes (por ahora, solo *sexo*) valen cero. En nuestro ejemplo, $e^{-0.94} = 0,39$ es la *odds* del suceso consumir marihuana entre las mujeres: el número de mujeres que consumen marihuana es un 39% del número de mujeres que no la consumen.
- Coeficiente $\hat{\lambda}_{\text{hombres}, st}$. El valor exponencial de este coeficiente es la *odds ratio* que compara las *odds* de los grupos definidos por la variable independiente. En nuestro ejemplo, $e^{1.69} = 5,42$ es la *odds ratio* que compara la *odds* los hombres (2,12) con la *odds* de las mujeres (0,39). Por tanto, la *odds* de consumir marihuana entre los hombres es 5,42 veces mayor que entre las mujeres.

Más de una variable independiente

Consideremos ahora un diseño con tres variables (X , Y y Z) en el que X e Y son variables categóricas que se toman como *independientes* y Z es una variable dicotómica que se toma como *dependiente*.

Si Z es una variable dicotómica ($k = 1, 2$), la probabilidad condicional de que Z tome el valor $k = 1$ en las diferentes categorías de X e Y viene dada por

$$P(Z=1) = \pi_{ij1} / (\pi_{ij1} + \pi_{ij2}) = \pi_{ij1} / \pi_{ij+} \quad [8.40]$$

Tomando frecuencias absolutas en lugar de relativas y desarrollando, el *logit* de Z en cada combinación de niveles de las variables X e Y puede formularse como

$$\text{logit}(Z=1) = \log_e(m_{ij1}/m_{ij2}) \quad [8.41]$$

De nuevo, lo que se está intentando modelar no es el contenido de las casillas, m_{ijk} , sino el cociente entre las frecuencias correspondientes a las dos categorías de la variable dependiente. Unas sencillas transformaciones llevan a

$$\begin{aligned} \log_e(m_{ij1}/m_{ij2}) = & [\lambda_{Z(1)} - \lambda_{Z(2)}] + [\lambda_{XZ(i1)} - \lambda_{XZ(i2)}] + \\ & + [\lambda_{YZ(j1)} - \lambda_{YZ(j2)}] + [\lambda_{XYZ(ij1)} - \lambda_{XYZ(ij2)}] \end{aligned} \quad [8.42]$$

Utilizando ahora una notación similar a la de los modelos loglineales, el modelo que incluye todos los términos posibles (es decir, los correspondientes a las dos variables independientes individualmente consideradas y el correspondiente a la interacción entre ambas) puede expresarse como:

$$\text{logit}(Z=1) = \log_e(m_{ij1}/m_{ij2}) = \xi + \xi_{X(i)} + \xi_{Y(j)} + \xi_{XY(ij)} \quad [8.43]$$

donde $\xi = \lambda_{Z(1)} - \lambda_{Z(2)}$ (término constante de la ecuación), recoge la presencia de la variable dependiente Z ; $\xi_{X(i)} = \lambda_{XZ(i1)} - \lambda_{XZ(i2)}$ representa la asociación parcial entre X y Z ; $\xi_{Y(j)} = \lambda_{YZ(j1)} - \lambda_{YZ(j2)}$ representa la asociación parcial entre Y y Z ; y, finalmente, $\xi_{XY(ij)} = \lambda_{XYZ(ij1)} - \lambda_{XYZ(ij2)}$ representa la asociación triple entre X , Y y Z .

Todos los términos λ_h suman cero para cada efecto (pues se conciben como desviaciones de algún promedio):

$$\sum_k \lambda_{Z(k)} = \sum_k \lambda_{XZ(ik)} = \sum_k \lambda_{YZ(jk)} = \sum_k \lambda_{XYZ(ijk)} = 0 \quad [8.44]$$

Teniendo esto en cuenta y que la variable Z es dicotómica, los parámetros $\lambda_{Z(2)}$, $\lambda_{XZ(i2)}$, $\lambda_{YZ(j1)}$ y $\lambda_{XYZ(ij2)}$ son redundantes. Y como el SPSS estima los parámetros de un modelo logit fijando en cero los parámetros redundantes, se tiene:

$$\begin{aligned} \xi &= \lambda_{Z(1)} - \lambda_{Z(2)} = \lambda_{Z(1)} \\ \xi_{X(i)} &= \lambda_{XZ(i1)} - \lambda_{XZ(i2)} = \lambda_{XZ(i1)} \\ \xi_{Y(j)} &= \lambda_{YZ(j1)} - \lambda_{YZ(j2)} = \lambda_{YZ(j1)} \\ \xi_{XY(ij)} &= \lambda_{XYZ(ij1)} - \lambda_{XYZ(ij2)} = \lambda_{XYZ(ij1)} \end{aligned} \quad [8.45]$$

Por supuesto, también es posible formular un modelo logit que proponga efectos aditivos de las variables X e Y sobre la variable Z (es decir, un modelo sin la interacción XY). Para ello, basta con eliminar de [8.43] el término referido a la interacción triple.

Correspondencia entre los modelos logit y los loglineales

Existen diversos procedimientos para ajustar modelos logit del tipo descrito en los apartados anteriores (ver por ejemplo, Cox, 1970; Haberman, 1978; Theil, 1970; etc.). No obstante, Bishop (1969) y Goodman (1971) han demostrado que los modelos loglineales pueden adaptarse para ofrecer resultados equivalentes a los que se obtienen con un modelo logit: *un modelo logit es equivalente al modelo loglineal que incluye los términos correspondientes a las interacciones de las que forma parte la variable dependiente más los términos correspondientes a todas las interacciones entre las variables independientes que forman parte de alguna interacción con la variable dependiente*.

Por ejemplo, en una situación con tres variables (X , Y y Z , con Z dependiente), el modelo logit que incluye los efectos principales de X e Y pero no la interacción XY es equivalente al modelo loglineal $[XY, XZ, YZ]$. En una situación con cuatro variables (X , Y , Z y V , con V dependiente), el modelo logit que incluye los efectos principales X , Y y Z , y la interacción YZ es equivalente al modelo loglineal $[XYZ, XV, YZV]$. Y el modelo logit que incluye los efectos principales de X , Y y Z , y las interacciones XY e YZ es equivalente al modelo loglineal $[XYZ, XYV, YZV]$.

Antes de explicar el procedimiento **Logit** vamos a ver cómo utilizar el procedimiento **Loglineal > Selección de modelo** para obtener un modelo logit a partir del ajuste de modelos loglineales jerárquicos. Vamos a utilizar los datos de la Tabla 8.37, que se refieren

a una muestra de 200 sujetos clasificados a partir de tres variables categóricas. Queremos averiguar si las variables independientes X e Y están relacionadas con la variable dependiente Z ; es decir, queremos encontrar el modelo logit que mejor se ajusta a estos datos cuando la variable Z se toma como variable dependiente. Para esto, lo que vamos a hacer es buscar el mejor modelo loglineal.

La forma de buscar el modelo loglineal que mejor se ajusta a estos datos consiste en proceder por pasos (ya hemos hecho esto en el apartado *Ajuste por pasos*), comparando modelos alternativos que difieran en un solo término hasta encontrar el modelo capaz de ofrecer el mejor ajuste con el menor número de términos. Para aplicar esta estrategia por pasos:

- Reproducir en el *Editor de datos* los datos de la Tabla 8.37 tal como muestra la Figura 8.8 y ponderar los casos del archivo con la variable *ncasos* mediante la opción **Ponderar casos** del menú **Datos** (en el ejemplo de la Figura 8.8, las variables de la

Tabla 8.37. Frecuencias obtenidas al clasificar una muestra de 100 sujetos en X = “concepción que se tiene de la inteligencia”; Y = “tipo de mensajes autodirigidos al realizar una tarea de rendimiento”; y Z = “tipo de meta motivacional hacia la que se orienta la conducta en esa misma tarea”

(X) <i>Concepción inteligencia</i>	(Y) <i>Tipo de automensajes</i>	(Z) <i>Tipo de meta</i>		<i>Totales XY</i>	<i>Totales X</i>
		<i>Aprendizaje</i>	<i>Ejecución</i>		
<i>Destreza</i>	<i>Instrumentales</i>	43	11	54	98
	<i>Atribucionales</i>	21	10	31	
	<i>Otros</i>	4	9	13	
<i>Rasgo</i>	<i>Instrumentales</i>	19	28	47	102
	<i>Atribucionales</i>	7	38	45	
	<i>Otros</i>	2	8	10	
<i>Totales de Z</i>		96	104		200

Figura 8.8. Datos de la Tabla 8.37 reproducidos en el *Editor de datos*

	intelig	automen	meta	ncasos
1	1	1	1	43
2	1	1	2	11
3	1	2	1	21
4	1	2	2	10
5	1	3	1	4
6	1	3	2	9
7	2	1	1	19
8	2	1	2	28
9	2	2	1	7
10	2	2	2	38
11	2	3	1	2
12	2	3	2	8

Tabla 8.37 se han nombrado *intelig* (X), *automen* (Y) y *meta* (M); la variable *nca*-*casos* recoge las frecuencias de las casillas; estos datos están disponibles en el archivo *Loglineal logit*, el cual puede descargarse de la página web del manual).

- Seleccionar la opción **Loglineal > Selección de modelo** del menú **Analizar** para acceder al cuadro de diálogo *Análisis loglineal: Selección de modelo* y trasladar las variables *intelig*, *automen* y *meta* a la lista **Factores**.
- Pulsar el botón **Definir rango** para asignar a cada variable seleccionada el correspondiente rango de códigos: 1 y 2 para las variables *intelig* y *meta*; 1 y 3 para la variable *automen*.

Aceptando estas elecciones el *Visor* ofrece, entre otros, los resultados que muestra la Tabla 8.38. La tabla ofrece los resultados del proceso de *eliminación hacia atrás* partiendo del modelo saturado. En el último paso (*Paso 3* en el ejemplo) se indica que el modelo final es el que incluye las interacciones *intelig* $\times$ *meta* y *automen* $\times$ *meta*. Por tanto, y de acuerdo con el principio de jerarquía, el modelo loglineal final es el modelo que incluye los términos referidos a esas dos interacciones dobles más los referidos a los efectos principales contenidos en ellas, es decir, el modelo $[XZ, YZ]$:

$$\log_e(m_{ijk}) = \lambda + \lambda_{\text{intelig}(i)} + \lambda_{\text{automen}(j)} + \lambda_{\text{meta}(k)} + \lambda_{\text{intelig} \times \text{meta}(ik)} + \lambda_{\text{automen} \times \text{meta}(jk)}$$

Y el modelo loglineal que equivale al modelo logit que estamos buscando se obtiene añadiendo la interacción entre las variables independientes que están relacionadas con la dependiente; en nuestro ejemplo, la interacción XY (pues tanto X como Y interaccionan con Z). Se llega así al modelo de *asociación homogénea* $[XY, XZ, YZ]$, es decir, a un modelo que asume que la relación entre cada par de variables es la misma independientemente del nivel de la tercera variable que se considere.

Y, teniendo en cuenta la correspondencia existente entre los modelos logit y los loglineales, el modelo logit equivalente al loglineal encontrado es

$$\text{logit}(meta = 1) = \log_e(m_{ij(\text{aprendizaje})} / m_{ij(\text{ejecución})}) = \xi + \xi_{\text{intelig}(i)} + \xi_{\text{automen}(j)}$$

Tabla 8.38. Pasos del proceso de eliminación hacia atrás

Paso	Efectos	Chi-cuadrado	gl	Sig.
0 Clase generadora	intelig*automen*meta	,00	0	.
Efecto eliminado 1	intelig*automen*meta	2,64	2	,267
1 Clase generadora	intelig*automen, intelig*meta, automen*meta	2,64	2	,267
Efecto eliminado 1	intelig*automen	3,64	2	,162
2	intelig*meta	36,57	1	,000
3	automen*meta	16,02	2	,000
2 Clase generadora	intelig*meta, automen*meta	6,28	4	,179
Efecto eliminado 1	intelig*meta	36,32	1	,000
2	automen*meta	15,77	2	,000
3 Clase generadora	intelig*meta, automen*meta	6,28	4	,179

Este modelo logit incluye los efectos principales de las dos variables independientes pero no el efecto de la interacción. Por tanto, se puede concluir que la orientación motivacional (*meta*) está relacionada tanto con la concepción que se tiene de la inteligencia (*intelig*) como con los mensajes que los sujetos se autodirigen al realizar una tarea de rendimiento (*automen*), pero no con la combinación de ambas cosas. El significado de este modelo se puede precisar construyendo dos tablas bidimensionales:

- Seleccionar la opción **Estadísticos descriptivos > Tablas de contingencias** del menú **Analizar** para acceder al cuadro de diálogo *Tablas de contingencias*.
- Trasladar las variables *intelig* y *automen* a la lista **Filas** y la variable *meta* a la lista **Columnas**.
- Pulsar el botón **Casillas** para acceder al subcuadro de diálogo *Tablas de contingencias: Mostrar en las casillas* y marcar la opción **Tipificados corregidos** del recuadro **Residuos**.

Aceptando estas elecciones el *Visor* ofrece los resultados que muestran las Tablas 8.39 y 8.40. Ambas incluyen los residuos tipificados corregidos (calculados asumiendo independencia entre las variables). Recordemos que, puesto que estos residuos se distribuyen de forma aproximadamente normal, con media 0 y desviación típica 1, los valores muy grandes en valor absoluto (mayores que 1,96 si se utiliza un nivel de confianza de 0,95) delatan casillas con más casos (si el residuo es positivo) o menos (si el residuo es negativo) de los que cabría esperar si realmente las variables fueran independientes.

Los residuos de la Tabla 8.39 indican que entre los sujetos que conciben la inteligencia como una *destreza* se da un desplazamiento significativo de casos desde la categoría *metas de ejecución* (−5,9) hacia la categoría *metas de aprendizaje* (5,9); mientras

Tabla 8.39. Tabla de contingencias de *intelig* por *meta*

			meta	
			aprendizaje	ejecución
intelig	Destreza	Recuento	68	30
		Residuos corregidos	5,9	-5,9
	Rasgo	Recuento	28	74
		Residuos corregidos	-5,9	5,9

Tabla 8.40. Tabla de contingencias de *automen* por *meta*

			meta	
			aprendizaje	ejecución
automen	Instrum	Recuento	62	39
		Residuos corregidos	3,8	-3,8
	Atribuc	Recuento	28	48
		Residuos corregidos	-2,5	2,5
	Otras	Recuento	6	17
		Residuos corregidos	-2,2	2,2

que entre los sujetos que conciben la inteligencia como un *rasgo* se da un desplazamiento significativo de casos desde la categoría metas de *aprendizaje* (-5,9) hacia la categoría metas de *ejecución* (5,9).

Los residuos tipificados corregidos de la Tabla 8.40 indican que entre los sujetos que se dirigen automensajes *instrumentales* se da un desplazamiento significativo de casos desde la categoría metas de *ejecución* (-3,8) hacia la categoría metas de *aprendizaje* (3,8); mientras que entre los sujetos que se dirigen automensajes *atribucionales* se da un desplazamiento significativo de casos desde la categoría metas de *aprendizaje* (-2,5) hacia la categoría metas de *ejecución* (2,5).

Combinando ambas pautas de asociación puede concluirse que los sujetos que conciben la inteligencia como una *destreza* y que se dirigen automensajes *instrumentales* tienden a manifestar metas motivacionales de *aprendizaje* y a no manifestar metas motivacionales de *ejecución* (son sujetos más preocupados por aprender de la tarea que por el resultado de la misma), mientras que los sujetos que conciben la inteligencia como un *rasgo* y que se dirigen automensajes *atribucionales* tienden a manifestar metas motivacionales de *ejecución* y a no manifestar metas de *aprendizaje* (son sujetos más preocupados por el resultado de la tarea que por aprender de ella).

El procedimiento **Logit**

Aunque es posible llegar a un modelo logit a partir del ajuste de modelos loglineales, también es posible utilizar el procedimiento **Logit** para ajustar modelos logit concretos. Este procedimiento no permite el ajuste por pasos, pero incluye otras prestaciones; por ejemplo, estima los parámetros del modelo elegido y calcula varios tipos de residuos (incluyendo los tipificados corregidos) y algunas medidas del tamaño del efecto.

Puesto que el procedimiento **Logit** no permite ajustar modelos por pasos, antes de utilizarlo es necesario tener una idea acerca del modelo concreto que se desea ajustar. Si no se tiene una idea clara sobre esto, una buena forma de proceder consiste en aplicar el procedimiento **Selección de modelo** para encontrar el mejor modelo loglineal y, a continuación, utilizar el procedimiento **Logit** para obtener las estimaciones y residuos de ese modelo. En el ejemplo del apartado anterior hemos llegado a la conclusión de que el modelo logit que ofrece el mejor ajuste con el menor número de parámetros es el que incluye los efectos individuales de las variables *intelig* y *automen*. Ahora vamos a utilizar el procedimiento **Logit** para obtener información adicional sobre ese modelo:

- Seleccionar la opción **Loglineal > Logit** del menú **Analizar** para acceder al cuadro de diálogo *Análisis loglineal logit* y trasladar la variable *meta* a la lista **Dependiente** y las variables *intelig* y *automen* a la lista **Factores**¹².

¹² Generalmente interesará trabajar con variables dependientes dicotómicas, pero el procedimiento admite variables politómicas. La lista **Covariables de casilla** admite variables independientes cuantitativas. Cuando se utiliza una covariable cuantitativa, el SPSS no utiliza los valores individuales de cada caso, sino la media de cada casilla. Las listas **Estructura de las casillas** y **Variables de contraste** tienen la misma utilidad que en el procedimiento **General** (ver más atrás, en este mismo capítulo, los apartados *Estructura de las casillas* y *Comparaciones entre niveles*).

- Pulsar el botón **Modelo** para acceder al subcuadro de diálogo *Análisis loglineal logit: Modelo*, marcar la opción **Personalizado**¹³ y trasladar las variables *intelig* y *automen* a la lista **Términos del modelo** vigilando que en el menú desplegable **Construir términos** esté seleccionada la opción **Efectos principales**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones se obtienen entre otros, los resultados que muestran las Tablas 8.41 a 8.45.

La Tabla 8.41 incluye información sobre el archivo: se están utilizando 12 casos válidos (que en realidad son 200 tras la ponderación) y no se ha desechado ningún caso por tener valor perdido; la tabla consta de 12 casillas sin ceros estructurales ni muestrales (12 patrones de variabilidad); y se han incluido en el análisis tres variables: *meta* (con 2 categorías), *intelig* (con 2 categorías) y *automen* (con 3 categorías).

Tabla 8.41. Información sobre los datos

		N
Casos	Válidos	12
	Perdidos	0
	Válidos ponderados	200
Casillas	Casillas definidas	12
	Ceros estructurales	0
	Ceros de muestreo	0
Categorías	meta	2
	intelig	2
	automen	3

Ajuste global: significación estadística

La Tabla 8.42 contiene los dos estadísticos de bondad ajuste¹⁴: la razón de verosimilitudes (2,64) y el estadístico de Pearson (2,95). Ambos estadísticos aparecen acompañados de sus correspondientes grados de libertad (*gl*) y niveles críticos (*sig.*); puesto que en ambos casos el nivel crítico es mayor que 0,05, puede asumirse que el modelo propuesto ofrece un buen ajuste a los datos.

Las dos notas a pie de tabla recuerdan qué esquema de muestreo se está utilizando (*modelo: logit multinomial*; ver, en el Apéndice 8, el apartado *Esquemas de muestreo*) y qué términos concretos incluye el modelo *logit* que se está ajustando (*diseño*). El modelo logit del ejemplo incluye la variable dependiente *meta* y los efectos principales de las variables *intelig* y *automen* (además del término constante). El SPSS identifica estos

¹³ Al construir un modelo personalizado debe tenerse en cuenta que el procedimiento **Logit** no sigue el principio de jerarquía. Esto significa que en la lista **Construir términos** deben incluirse todos los términos que se desea que formen parte del modelo.

¹⁴ El valor de estos estadísticos es idéntico al que se obtiene al ajustar el modelo loglineal [XY, XZ, XY] con el procedimiento **Selección de modelo**.

efectos mediante las interacciones *meta* $\times$ *automen* (efecto principal de la variable *automen*) e *intelig* $\times$ *meta* (efecto principal de la variable *intelig*).

Tabla 8.42. Estadísticos de bondad de ajuste

	Valor ^{a,b}	gl	Sig.
Razón de verosimilitudes	2,64	2	,267
Chi-cuadrado de Pearson	2,95	2	,228

a. Modelo: Logit multinomial

b. Diseño: Constante + meta + meta * automen + meta * intelig

A continuación (Tabla 8.43) se ofrecen las frecuencias observadas y las esperadas (en valor absoluto y porcentual), los residuos en *bruto* o no tipificados, los residuos tipificados, los residuos tipificados corregidos y los residuos de desviación; todos estos valores corresponden al modelo logit que se está ajustando (o, si se prefiere, al modelo loglineal que equivale al modelo logit que se está ajustando). En coherencia con la conclusión ya adoptada de que el modelo ofrece un buen ajuste a los datos, todos los residuos tipificados corregidos y todos los residuos de desviación tienen valores comprendidos entre -1,96 y 1,96.

Tabla 8.43. Frecuencias y residuos

intelig	automen	meta	Observado		Esperado		Resid.	Resid. tipificad.	Resid. corregid.	Resid. desvian.
			n	%	n	%				
Destreza	Instrum	aprend.	43	79,6%	43,70	80,9%	-,70	-,24	-,47	-1,18
		ejecuc.	11	20,4%	10,30	19,1%	,70	,24	,47	1,21
	Atribuc	aprend.	21	67,7%	19,18	61,9%	1,82	,67	1,24	1,95
		ejecuc.	10	32,3%	11,82	38,1%	-1,82	-,67	-1,24	-1,83
	Otras	aprend.	4	30,8%	5,11	39,3%	-1,11	-,63	-1,44	-1,40
		ejecuc.	9	69,2%	7,89	60,7%	1,11	,63	1,44	1,54
Rasgo	Instrum	aprend.	19	40,4%	18,30	38,9%	,70	,21	,47	1,20
		ejecuc.	28	59,6%	28,70	61,1%	-,70	-,21	-,47	-1,18
	Atribuc	aprend.	7	15,6%	8,82	19,6%	-1,82	-,68	-1,24	-1,80
		ejecuc.	38	84,4%	36,18	80,4%	1,82	,68	1,24	1,93
	Otras	aprend.	2	20,0%	,89	8,9%	1,11	1,24	1,44	1,80
		ejecuc.	8	80,0%	9,11	91,1%	-1,11	-1,24	-1,44	-1,44

Entre los resultados que el procedimiento ofrece por defecto se incluyen algunos estadísticos que permiten estudiar el grado de asociación existente entre la variable dependiente y las independientes. En concreto, el procedimiento **Logit** ofrece los índices de **entropía** y de **concentración** (Tabla 8.44).

El procedimiento genera una tabla de dispersión similar a la tabla resumen de un ANOVA en un análisis de regresión lineal. La dispersión total de la variable dependiente (la diferencia existente entre las frecuencias marginales de la variable depen-

diente) se divide en dos partes: la dispersión explicada por el modelo (*modelo*) y la no explicada por el modelo (*residual*). Los estadísticos de entropía y concentración que ofrece la Tabla 8.44 pueden utilizarse para valorar si el modelo propuesto difiere del modelo nulo, es decir, del modelo que únicamente incluye la constante (ver Haberman, 1982).

Comenzando con la medida de *entropía*, si el modelo logit propuesto es correcto, el doble de la entropía debida al modelo se distribuye según ji-cuadrado con los mismos grados de libertad que el modelo logit propuesto. En nuestro ejemplo, $2(26,17)=52,34$ se distribuye según ji-cuadrado con 2 grados de libertad (este valor es la diferencia entre la razón de verosimilitudes del modelo que únicamente incluye la constante y la razón de verosimilitudes del modelo que se está ajustando). La probabilidad de encontrar un valor como 52,34 o mayor en la distribución ji-cuadrado con 2 grados de libertad toma un valor muy próximo a cero¹⁵; como este valor es menor que 0,05, puede afirmarse que el modelo logit que se está ajustando difiere significativamente del modelo nulo; o, lo que es lo mismo, que las variables independientes *intelig* y *automen* están relacionadas con la variable dependiente *meta*.

Con la medida de *concentración* se llega a la misma conclusión. El cociente entre la concentración debida al modelo dividida entre sus grados de libertad ($24,20/3=8,07$) y la no debida al modelo dividida entre sus grados de libertad ($75,64/196=0,39$) se distribuye según el modelo de probabilidad *F* con los grados de libertad del numerador y del denominador. La distribución del cociente $8,07/0,39=20,69$ se aproxima a la distribución *F* con 3 y 196 grados de libertad. La probabilidad de obtener valores iguales o mayores que 20,69 en una distribución *F* con 3 y 196 grados de libertad es muy próxima a cero¹⁶; por tanto, puede concluirse que el modelo logit que se está ajustando difiere significativamente del modelo nulo.

Tabla 8.44. Análisis de la dispersión

	Entropía	Concentración	gl
Modelo	26,17	24,20	3
Residual	112,30	75,64	196
Total	138,47	99,84	199

Ajuste global: significación sustantiva

Las medidas de concentración y entropía pueden interpretarse (ver Magidson, 1981) como medidas de la proporción de dispersión de la variable dependiente que es atribuible al modelo (de modo similar a como se interpreta R^2 en un análisis de regresión lineal). Cuando se utiliza la medida de entropía, el cociente entre la dispersión explicada por el

¹⁵ Esta probabilidad puede obtenerse con la función SIG.CHISQ de la opción **Calcular** del menú **Transformar**, utilizando como expresión numérica: SIG.CHISQ(52.34, 2).

¹⁶ Esta probabilidad puede obtenerse con la función SIG.F de la opción **Calcular** del menú **Transformar**, utilizando como expresión numérica: SIG.F(4.84,3,96).

modelo y la dispersión total ($26,17/138,47=0,19$) está indicando que el modelo consigue explicar un 19% de la dispersión de la variable dependiente. Si se utiliza la medida de concentración, el porcentaje de dispersión explicada sube al 24% (estos valores se ofrecen en la Tabla 8.45).

Con tablas bidimensionales, la medida de entropía coincide con el coeficiente de incertidumbre y la medida de concentración con el cuadrado del coeficiente *tau-b* de Kendal (ver Capítulo 3 del segundo volumen).

Debe tenerse en cuenta que, aunque estas medidas de asociación se interpretan de forma similar a como se interpreta el coeficiente de determinación R^2 en un modelo de regresión lineal, lo cierto es que su valor puede ser pequeño incluso cuando existe una fuerte asociación entre las variables involucradas.

Tabla 8.45. Medidas de asociación

Entropía	,19
Concentración	,24

Interpretación de los coeficientes de un modelo logit

El subcuadro de diálogo *Opciones* permite obtener información adicional y controlar algunos detalles del análisis. En concreto, permite obtener las estimaciones de los parámetros del modelo propuesto (opción **Estimaciones**). La Tabla 8.46 muestra las estimaciones correspondientes a los parámetros de nuestro ejemplo. Recuérdese que a los parámetros redundantes se les asigna un valor de cero (esta circunstancia se indica en una nota a pie de tabla). Por tanto, las categorías o combinación de categorías a las que les corresponde un parámetro redundante son justamente las que se toman como punto de referencia para la comparación.

Para cada parámetro no redundante la tabla ofrece su valor estimado, su error típico, su valor tipificado (Z ; se obtiene dividiendo el valor estimado entre su error típico) y un intervalo de confianza que permite decidir si el parámetro es distinto de cero: los intervalos cuyos límites no incluyen el valor cero corresponden a parámetros que deben estar en el modelo para que éste ofrezca un buen ajuste a los datos.

El modelo logit que estamos ajustando incluye 18 parámetros; de éstos, 8 son redundantes. De los 10 no redundantes, los 6 primeros (*constantes*) pueden ignorarse porque corresponden a los términos del modelo loglineal no incluidos en el correspondiente modelo logit. En nuestro ejemplo, el único término *loglineal* no incluido es la interacción *intelig* $\times$ *automen*; los términos *constantes* recogen las estimaciones correspondientes a esta interacción.

En consecuencia, el modelo logit que estamos ajustando solo incluye 4 parámetros no redundantes: la constante del modelo, que recoge el efecto de la variable dependiente *meta* (puesto que la variable *meta* solo tiene dos niveles, el segundo parámetro es redundante); los dos parámetros correspondientes al efecto de la variable independiente *automen* (puesto que la variable *meta* tiene dos niveles y la variable *automen* tiene tres,

Tabla 8.46. Estimaciones de los parámetros

Parámetro	Estim.	Error típico	Z	Sig.	Intervalo de confianza al 95%	
					L. inf.	L. sup.
Constante						
[intelig = 1] * [automen = 1]	2,332 ^a					
[intelig = 1] * [automen = 2]	2,470 ^a					
[intelig = 1] * [automen = 3]	2,065 ^a					
[intelig = 2] * [automen = 1]	3,357 ^a					
[intelig = 2] * [automen = 2]	3,589 ^a					
[intelig = 2] * [automen = 3]	2,210 ^a					
[meta = 1]	-2,330	,57	-4,08	,000	-3,45	-1,21
[meta = 2]	0 ^b	.	.	.	.	.
[meta = 1] * [automen = 1]	1,879	,57	3,31	,001	,77	2,99
[meta = 1] * [automen = 2]	,918	,58	1,59	,112	-,22	2,05
[meta = 1] * [automen = 3]	0 ^b	.	.	.	.	.
[meta = 2] * [automen = 1]	0 ^b	.	.	.	.	.
[meta = 2] * [automen = 2]	0 ^b	.	.	.	.	.
[meta = 2] * [automen = 3]	0 ^b	.	.	.	.	.
[meta = 1] * [intelig = 1]	1,896	,33	5,69	,000	1,24	2,55
[meta = 1] * [intelig = 2]	0 ^b	.	.	.	.	.
[meta = 2] * [intelig = 1]	0 ^b	.	.	.	.	.
[meta = 2] * [intelig = 2]	0 ^b	.	.	.	.	.

a. Las constantes no son parámetros bajo el supuesto multinomial. Por tanto, no se calculan sus errores típicos.

b. Este parámetro se ha definido como cero ya que es redundante.

los parámetros correspondientes al segundo nivel de la variable *meta* y al tercer nivel de la variable *automen* son redundantes); y el correspondiente al efecto de la variable *intelig* (puesto que tanto la variable *meta* como la variable *intelig* tienen solo dos niveles, los parámetros correspondientes a los segundos niveles de ambas variables son redundantes).

El valor exponencial de la constante [meta = 1] es una *odds*: la *odds* de la primera categoría de la variable dependiente respecto de la segunda cuando todas las variables independientes valen cero. Por tanto, $e^{-2,33} = 0,097$ es una estimación de la frecuencia con la que se dan metas de aprendizaje en comparación con la frecuencia con la que se dan metas de ejecución, pero solo para quienes conciben la inteligencia como un *rasgo* y se dirigen *otros* automensajes (es decir, para quienes pertenecen a las categorías de las variables independientes cuyos parámetros se han fijado en cero).

Cada una de las restantes estimaciones que ofrece la Tabla 8.46 (tres en total) es una *odds ratio* con un significado concreto relacionado con el efecto de las variables incluidas en el análisis. Para facilitar la interpretación de estas estimaciones las hemos ordenado tal como muestra la Tabla 8.47.

- Efecto de la *inteligencia*. El único parámetro no redundante asociado a la variable *intelig* tiene un valor estimado de 1,896. El signo positivo del coeficiente indica que, entre quienes conciben la inteligencia como una *destreza* es más probable encontrar metas de *aprendizaje* que entre quienes la conciben como un *rasgo*. El va-

lor exponencial del coeficiente ($e^{1,896} = 6,66$) indica que la *odds* de las metas de *aprendizaje* es 6,66 veces mayor entre quienes conciben la inteligencia como una *destreza* que entre quienes la conciben como un *rasgo* (cualquiera que sea el tipo de automensajes que utilicen).

- Efecto de los *automensajes*. El primer parámetro no redundante asociado a la variable *automen* tiene un valor estimado de 1,879 y corresponde a la categoría *instrumentales*. Su valor exponencial $e^{1,879} = 6,55$ indica que la *odds* de las metas de *aprendizaje* es 6,55 veces mayor entre quienes se dirigen automensajes *instrumentales* que entre quienes se dirigen *otros* automensajes (la categoría *otros* es la categoría de referencia para la comparación porque es la categoría cuyo parámetro se ha fijado en cero).

El segundo parámetro no redundante asociado a la variable *automen* tiene un valor estimado de 0,918 y corresponde a la categoría *atribucionales*. Su valor exponencial $e^{0,918} = 2,50$ indica que la *odds* de las metas de *aprendizaje* es 2,50 veces mayor entre quienes se dirigen automensajes *atribucionales* que entre quienes se dirigen *otros* automensajes. No obstante, esta diferencia no alcanza la significación estadística (*sig.* = 0,112).

Tabla 8.47. Estimaciones de los parámetros (valores exponenciales entre paréntesis)

<i>Meta</i>	<i>Inteligencia</i>		<i>Automensajes</i>		
	Destreza	Rasgo	Instrum.	Atribuc.	Otros
Aprendizaje	1,896 (6,66)	0 (1)	1,879 (6,55)	0,918 (2,50)	0 (1)
Ejecución	0 (1)	0 (1)	0 (1)	0 (1)	0 (1)

Las estimaciones de la Tabla 8.46 permiten obtener los pronósticos del modelo. Y con los pronósticos ya es posible calcular cualquier *odds ratio*, incluidas las que acabamos de interpretar en los párrafos anteriores.

Recordemos que el modelo logit que estamos ajustando (el modelo de independencia con dos variables independientes, es decir, el modelo que incluye el efecto de ambas variables independientes pero no su interacción) adopta la forma:

$$\text{logit}(meta = 1) = \log_e(m_{ij1}/m_{ij2}) = \xi + \xi_{X(i)} + \xi_{Y(j)}$$

con $\xi = \lambda_{Z(1)}$, $\xi_{X(i)} = \lambda_{XZ(i1)}$ y $\xi_{Y(j)} = \lambda_{YZ(j1)}$. Por tanto,

$$\text{logit}(meta = 1) = \log_e(m_{ij1}/m_{ij2}) = \lambda_{Z(1)} + \lambda_{XZ(i1)} + \lambda_{YZ(j1)}$$

Tenemos X = “intelig”, Y = “automen”. Además, Z = “meta” (con 1 = “aprendizaje”). En consecuencia:

$$\text{logit}(meta = 1) = \lambda_{\text{aprend}} + \lambda_{\text{intelig} \times \text{meta} (i1)} + \lambda_{\text{automen} \times \text{meta} (j1)}$$

Por ejemplo, los pronósticos que ofrece el modelo para los sujetos que conciben la inteligencia como una *destreza* y se dirigen automensajes *instrumentales*, vienen dados por

$$\text{logit}(meta = 1) = \lambda_{\text{aprend}} + \lambda_{\text{destreza, aprend}} + \lambda_{\text{instrum, aprend}} = -2,33 + 1,896 + 1,879$$

Siguiendo esta lógica podemos obtener los seis pronósticos que ofrece el modelo para las seis casillas resultantes de combinar las dos categorías de la variable *intelig* con las tres de la variable *automen*:

1. $\text{logit}(meta = 1 | \text{destreza, instrum.}) = -2,33 + 1,896 + 1,879 = 1,445 \text{ (4,24)}$
 $\text{logit}(meta = 1 | \text{rasgo, instrum.}) = -2,33 + 0 + 1,879 = -0,451 \text{ (0,64)}$
2. $\text{logit}(meta = 1 | \text{destreza, atribuc.}) = -2,33 + 1,896 + 0,918 = 0,484 \text{ (1,62)}$
 $\text{logit}(meta = 1 | \text{rasgo, atribuc.}) = -2,33 + 0 + 0,918 = -1,412 \text{ (0,24)}$
3. $\text{logit}(meta = 1 | \text{destreza, otras}) = -2,33 + 1,896 + 0 = -0,434 \text{ (0,65)}$
 $\text{logit}(meta = 1 | \text{rasgo, otras}) = -2,33 + 0 + 0 = -2,33 \text{ (0,10)}$

Detrás de cada pronóstico en escala logit se ofrece, entre paréntesis, su valor exponencial. Estos pronósticos permiten apreciar varias cosas. Por ejemplo, el valor estimado para el término constante (-2,33) es efectivamente el pronóstico que ofrece el modelo cuando todas las variables independientes valen cero (rasgo, otros). Y el cociente entre los valores exponenciales de cada par de pronósticos *destreza-rasgo* es constante en cada categoría de la tercera variable (*automen*), lo cual ya sabemos que es así porque el modelo que hemos ajustado no incluye la interacción entre las variables independientes *intelig* y *automen*: salvo por pequeños detalles de redondeo, este valor constante es $4,24/0,64 = 1,62/0,24 = 0,65/0,10 = 6,66$, que no es otra cosa que el valor exponencial del coeficiente estimado para la variable *intelig* ($e^{1,896} = 6,66$) y que ya hemos interpretado señalando que la *odds* de las metas de *aprendizaje* es 6,66 veces mayor entre quienes conciben la inteligencia como una *destreza* que entre quienes la conciben como un *rasgo* (cualquiera que sea el tipo de automensajes que se utilicen).

Apéndice 8

Esquemas de muestreo

Para obtener las frecuencias de una tabla de contingencias pueden seguirse diferentes estrategias de recogida de datos. Estas estrategias, denominadas *esquemas de muestreo*, determinan las *distribuciones muestrales* de las frecuencias con las que se va a trabajar. Cada frecuencia de una tabla de contingencias es una variable aleatoria. Como tal, tiene su propia función de proba-

bilidad. Y esa función de probabilidad viene determinada por el tipo de muestreo utilizado. Este apartado incluye una breve exposición de los tres esquemas de muestreo más utilizados para describir las frecuencias (variables) de una tabla de contingencias: multinomial, multinomial condicional y Poisson.

Esquema multinomial

Quizá el más tradicional de estos procedimientos sea el esquema de muestreo **multinomial**. Este esquema es apropiado cuando lo que se pretende es (1) seleccionar de una población de interés una muestra aleatoria de tamaño n y (2) clasificar cada elemento (cada uno independientemente de cada otro) en las características definidas por las variables subyacentes.

Tomando como ejemplo los datos de la tabla 8.48, el esquema de muestreo multinomial habría llevado a seleccionar una muestra aleatoria de tamaño $n = 200$ y a clasificar a cada sujeto como *hombre-fumador*, *hombre-no fumador*, ..., *mujer-fumadora*, *mujer-no fumadora*, etc. En este escenario, las frecuencias observadas de una tabla bidimensional constituyen una variable aleatoria (resultado de la clasificación independiente de n observaciones aleatorias) con función de probabilidad:

$$\frac{n!}{\prod_i \prod_j n_{ij}!} \prod_i \prod_j \pi_{ij}^{n_{ij}} \quad [8.46]$$

donde π_{ij} representa la probabilidad de que un elemento aleatoriamente seleccionado pertenezca a la casilla ij . Y si la distribución de las frecuencias sigue el modelo de probabilidad multinomial, la distribución de cada casilla seguirá el modelo de probabilidad binomial con índice n y parámetro π_{ij} . De lo cual cabe deducir que el valor esperado (frecuencia esperada) de cada casilla, m_{ij} , vendrá dado por

$$E(n_{ij}) = m_{ij} = n\pi_{ij} \quad [8.47]$$

Bajo la hipótesis de independencia filas-columnas, las probabilidades teóricas π_{ij} pueden estimarse mediante

$$\hat{\pi}_{ij} = (n_{i+}/n)(n_{+j}/n) \quad [8.48]$$

Y una vez estimadas las probabilidades teóricas, ya es posible estimar las frecuencias esperadas. Así, la frecuencia esperada de, por ejemplo, la casilla (1, 1), es decir, de la casilla “hombre fumador”, puede estimarse mediante

$$\hat{m}_{11} = n\hat{\pi}_{11} = n(n_{1+}/n)(n_{+1}/n) = 200(94/200)(60/200) = 28,2$$

Tabla 8.48. Tabla de contingencias de sexo por *tabaquismo*

Sexo	Tabaquismo			Total
	Fumadores	Exfumadores	No fumadores	
Hombres	18	7	69	94
Mujeres	42	6	58	106
Total	60	13	127	200

Esquema multinomial condicional

Otra forma diferente de proceder consiste en utilizar el esquema de muestreo **multinomial condicional**, también llamado **producto-multinomial**. De acuerdo con este esquema comenzaríamos seleccionando, por ejemplo, una muestra aleatoria de n_{1+} hombres y otra de n_{2+} mujeres y continuaríamos clasificando a los sujetos de cada muestra como fumadores, exfumadores o no fumadores. Al proceder de esta manera, ya no solo se fija de antemano el tamaño total de la muestra, n , como en el esquema de muestreo *multinomial*, sino que también se fijan los totales marginales de las filas (es decir, las frecuencias marginales n_{i+}).

Al proceder de esta manera, las frecuencias observadas de cada fila constituyen una variable aleatoria que se distribuye según el modelo de probabilidad multinomial $M(n_{i+}, \pi_{j|i})$, donde $\pi_{j|i}$ se refiere a la probabilidad condicional de la columna j dada la fila i . En este escenario, la función de probabilidad para cada fila viene dada por

$$\frac{n_{i+}!}{\prod_j n_{ij}!} \prod_j \pi_{j|i}^{n_{ij}} \quad [8.49]$$

Multiplicando las funciones multinomiales de cada fila se obtiene la función de probabilidad de la tabla entera. Y si las frecuencias de cada fila se distribuyen según el modelo multinomial, la distribución de las frecuencias de cada casilla sigue el modelo de probabilidad binomial con índice n_{i+} y parámetro $\pi_{j|i}$. Consecuentemente, las frecuencias esperadas vendrán dadas por

$$E(n_{ij}) = m_{ij} = n_{i+} \pi_{j|i} \quad [8.50]$$

Ahora no tenemos una única población, como ocurre en el muestreo multinomial, sino I poblaciones (tantas como filas). La hipótesis de independencia entre las filas y las columnas es equivalente a la hipótesis de *homogeneidad* de las I poblaciones, es decir, a la hipótesis de que la distribución de las J columnas es la misma en las I filas. Bajo esta hipótesis, la probabilidad de una casilla cualquiera, $\pi_{j|i}$, es la probabilidad condicional de la columna j dada la fila i . Asumiendo que las I filas son homogéneas, es decir, asumiendo que las I probabilidades condicionales $\pi_{j|i}$ de cada columna son iguales, es posible estimar todas ellas (todas las probabilidades de la misma columna) mediante un único valor: $\hat{\pi}_{j|i} = n_{+j}/n$. Y estimadas las probabilidades condicionales, ya es posible utilizar [8.50] para obtener las frecuencias esperadas. Así, la frecuencia esperada m_{11} , es decir, la frecuencia esperada de la casilla *hombre fumador* puede estimarse mediante

$$\hat{m}_{11} = n_{1+} \hat{\pi}_{1|1} = n_{1+} (n_{+1}/n) = 94 (60/200) = 28,2$$

Por supuesto, en lugar de fijar los totales de las filas podrían fijarse los totales de las columnas; en ese caso, el esquema de muestreo seguiría siendo el *multinomial condicional*, pero con n_{+j} fijo en lugar de n_{i+} .

Esquema de Poisson

El modelo de probabilidad de **Poisson** proporciona un tercer método o esquema de muestreo. De acuerdo con este método procederíamos sin establecer de antemano ni el n total ni los totales marginales. Nos limitaríamos a observar a los sujetos de una determinada población durante un periodo de tiempo establecido y a clasificarlos, independientemente unos de otros, según las variables de interés. Cuando se procede de esta manera, las frecuencias observadas constituyen una

variable aleatoria que se distribuye según el modelo de probabilidad de Poisson, por lo que la probabilidad correspondiente a cada casilla viene dada por

$$e^{-m_{ij}} \frac{m_{ij}^{n_{ij}}}{n_{ij}!} \quad [8.51]$$

Puesto que lo que ocurre en una casilla es independiente de lo que ocurre en cualquier otra (las observaciones son aleatoriamente seleccionadas y la asignación se hace independientemente para cada casilla), la función de probabilidad para la tabla entera vendrá dada por el producto de las IJ probabilidades [8.51].

Bajo la hipótesis de independencia entre las filas y las columnas, las frecuencias esperadas se obtienen, al igual que en el esquema multinomial, mediante $E(n_{ij}) = m_{ij} = n\pi_{ij}$. Por tanto, las estimaciones que se obtienen con este esquema de muestreo y con el multinomial son exactamente las mismas¹⁷.

Existen otros esquemas de muestreo (hipergeométrico, multinomial negativo, etc.) que también pueden servir para dar cuenta de las frecuencias de una tabla de contingencias. No obstante, los tres esquemas descritos, no solo son los más frecuentemente utilizados, sino que poseen una doble ventaja: permiten utilizar los mismos métodos inferenciales (por ejemplo, estimadores de máxima verosimilitud) y conducen a las mismas estimaciones para las frecuencias esperadas de una tabla de contingencias.

Estadísticos mínimo-suficientes

Un grupo de estadísticos es *suficiente* si permite reducir los datos de la tabla original y todavía es posible efectuar estimaciones sin perder información. Con un estadístico mínimo-suficiente esa reducción de datos es máxima: permite ignorar la parte de la tabla que contiene información redundante para la estimación.

En un modelo loglineal concreto, estos estadísticos mínimo-suficientes son las distribuciones marginales correspondientes a cada una de las configuraciones presentes en el símbolo del modelo. Para identificar esas distribuciones marginales:

1. Seleccionar los totales marginales correspondientes a los términos λ de mayor orden incluidos en el modelo.
2. Repetir el paso 1 para los términos λ de siguiente orden.
3. Eliminar cualquier total marginal redundante (si se conoce el total n_{ij+} para todos los valores de i y j , no es necesario conocer, por ejemplo, el total n_{i++} , pues éste puede obtenerse sumando los j totales n_{ij+} correspondientes a cada nivel de i).
4. Repetir los 3 pasos anteriores hasta revisar todos los términos λ .

Para ilustrar estos cuatro pasos, veamos cómo identificar los estadísticos-mínimo suficientes que corresponden al siguiente modelo loglineal:

$$\begin{aligned} \log_e(m_{ijk}) = & \lambda + \lambda_{X(i)} + \lambda_{Y(j)} + \lambda_{Z(k)} + \lambda_{Y(i)} \\ & + \lambda_{XY(ij)} + \lambda_{XZ(ik)} + \lambda_{YZ(jk)} + \lambda_{YV(jl)} + \lambda_{ZV(kl)} + \lambda_{XYZ(ijk)} \end{aligned}$$

¹⁷ En Bishop, Fienberg y Holland (1975, Capítulo 13) puede encontrarse un estudio detallado de estas y otras distribuciones de probabilidad.

Aplicando el paso 1 se obtiene n_{ijk+} , que es el marginal que corresponde al término de mayor orden presente en el modelo ($\lambda_{XYZ(ijk)}$). En el paso 2 se eligen los totales correspondientes a los términos de siguiente orden n_{ij++} , n_{i+k+} , n_{+jk+} , n_{+j+l} y n_{++kl} . Siguiendo con el paso 3 se deben eliminar los totales n_{ij++} , n_{i+k+} y n_{+jk+} , pues la información que contienen estos totales puede obtenerse a partir de n_{ijk+} . Solo quedan, por tanto, los totales n_{ij+k+} , n_{+j+l} y n_{++kl} . Estos son los estadísticos mínimo-suficientes, pues repitiendo los pasos 1 a 3 ya solo se obtienen totales marginales redundantes, en concreto: n_{i+++} , n_{+j++} y n_{+++k} (los cuales están contenidos en n_{ij+k+}) y n_{++++l} (que está contenido en, por ejemplo, n_{+j+l}).

Una vez identificados los estadísticos mínimo-suficientes correspondientes a un modelo dado, ya es posible utilizar el método de máxima verosimilitud para efectuar estimaciones.

Grados de libertad en un modelo loglineal

Los grados de libertad de la distribución de la razón de verosimilitudes se obtienen restando al número de patrones de variabilidad (el número de casillas de la tabla) el número de parámetros independientes estimados. La Tabla 8.49 contiene la información relativa a los modelos loglineales correspondientes a tablas de dos y tres dimensiones.

Tabla 8.49. Parámetros independientes y grados de libertad en modelos loglineales

Modelo	Parámetros independientes	Grados de libertad
[X, Y]	$I + J - 1$	$(I - 1)(J - 1)$
[XY]	IJ	0
[X, Y, Z]	$I + J + K - 2$	$IJK - I - J - K + 2$
[XY, Z]	$IJ + K - 1$	$(IJ - 1)(K - 1)$
[XZ, Y]	$IK + J - 1$	$(IK - 1)(J - 1)$
[YZ, X]	$JK + I - 1$	$(JK - 1)(I - 1)$
[XY, XZ]	$IJ + IK - I$	$I(J - 1)(K - 1)$
[XY, YZ]	$IJ + JK - J$	$J(I - 1)(K - 1)$
[XZ, YZ]	$IK + JK - K$	$K(I - 1)(J - 1)$
[XY, XZ, YZ]	$IJ + IK + JK - I - J - K + 1$	$(I - 1)(J - 1)(K - 1)$
[XYZ]	IJK	0

Por ejemplo, en el modelo de independencia correspondiente a una tabla de contingencias bidimensional se está estimando el término constante, los $I - 1$ parámetros independientes asociados a las I categorías de la variable X y los $J - 1$ parámetros independientes asociados a las J categorías de la variable Y . En total, $I + J - 1$ parámetros independientes. Puesto que la tabla tiene IJ casillas, los grados de libertad asociados a la razón de verosimilitudes del modelo de independencia son $IJ - (I + J - 1) = (I - 1)(J - 1)$.

Análisis de supervivencia

¿Cuánto tiempo sobrevive un paciente tras ser diagnosticado de una enfermedad terminal? ¿Cuál es la duración de los contratos de una determinada empresa? ¿Qué tiempo transcurre entre el inicio de un grado universitario y la obtención del título? Para responder a estas preguntas es necesario valorar el *tiempo transcurrido entre dos eventos*: el diagnóstico y la muerte, el contrato y el despido, la matriculación y la obtención del título. Y la respuesta no es trivial porque, en este tipo de situaciones, el evento que interesa estudiar (la muerte, el despido, la obtención del título) no necesariamente se da en todos los sujetos en el intervalo de tiempo en que se realiza el estudio.

El **análisis de supervivencia**, también llamado análisis de la *historia de eventos* y análisis de los *tiempos de espera*, incluye un conjunto de herramientas diseñadas para estudiar este tipo de datos. Se utiliza en campos como la epidemiología (para el estudio de la evolución de enfermedades y tratamientos), la sociología (para el estudio de cambios sociales, como el estado civil o la situación laboral), los seguros (para analizar el tiempo que permanecen los clientes con una póliza de riesgo), la ingeniería (para el estudio de la durabilidad de equipos y materiales), etc. Aunque no tiene por qué ser así, lo típico de este tipo de análisis es estudiar fenómenos que solo adoptan dos estados posibles: “vivo-muerto” o “recuperado-no recuperado” para pacientes, “funciona-no funciona” para máquinas, “estudia-abandona” para estudiantes, etc.

Quizá el análisis de supervivencia deba su nombre al hecho de que los primeros eventos que se estudiaron se referían a la muerte por enfermedad. Posiblemente también fue esto lo que llevó a llamar *terminal* al evento estudiado, si bien el evento no tiene por qué ser negativo: el *evento terminal* es un suceso, positivo o negativo, que los sujetos pueden experimentar en cualquier momento del estudio (la muerte, la recuperación, el despido, la obtención del título, etc.). La denominación de *terminal* no hace referencia a algo negativo, sino a su carácter irreversible: una vez que se produce, no

hay vuelta atrás; también hace referencia al hecho de que la observación o seguimiento de un sujeto concluye en el momento en que se produce el evento¹. El *evento terminal* es, junto con el *tiempo que tarda en aparecer*, el objetivo del análisis.

Tiempos de espera, eventos y casos censurados

En un análisis de supervivencia hay dos tipos de información (dos variables) imprescindibles: (1) la presencia o no del evento que se desea estudiar y (2) el tiempo que tarda en aparecer ese evento.

El primer paso del análisis consiste en definir el fenómeno de interés, el cual debe mostrar *dos estados posibles*: “está sucediendo” o “ha sucedido”. Ambos estados deben ser exclusivos y exhaustivos: un caso no puede adoptar ambos estados de manera simultánea y, en un momento dado, todos los casos deben adoptar uno de los dos estados. El **cambio de estado** indica que se ha producido el **evento** que se desea estudiar.

La presencia del evento se registra en una variable dicotómica cuyos valores reflejan los dos estados posibles (generalmente, 1 para el evento y 0 para el no-evento). Por ejemplo, al estudiar el tiempo de permanencia de sujetos drogodependientes en un programa de desintoxicación, el evento de interés podría ser el *abandono* antes de finalizar el programa; aquí, el fenómeno estudiado puede tomar dos valores: 1 = “abandona” y 0 = “sigue en tratamiento”.

Además de definir el evento es necesario registrar el momento exacto en el que aparece. Más concretamente, el *tiempo (t) transcurrido entre el inicio del seguimiento y la aparición del evento*. A este tiempo se le llama tiempo de **espera** o de **supervivencia** y es el dato característico de un análisis de supervivencia.

El problema que surge al analizar tiempos de espera es que, por lo general, el evento que interesa estudiar no siempre se produce en todos los sujetos que intervienen en el estudio. Un sujeto que no ha experimentado el evento al finalizar el seguimiento es un caso **censurado**. También se tiene un caso censurado cuando a un sujeto se le pierde la pista antes de finalizar el seguimiento (por ejemplo, porque muere accidentalmente antes de abandonar el tratamiento, porque continúa con el tratamiento en otro centro, etc.). En ambos casos se trata de sujetos de los que no se tiene constancia de que hayan experimentado el evento. La característica distintiva del análisis de supervivencia es que permite aprovechar la información relativa a los casos censurados: hasta donde se tiene noticia de ellos, al menos se sabe que todavía no han experimentado el evento².

El aspecto fundamental del análisis consiste en estudiar *el tiempo transcurrido entre el inicio del seguimiento y el momento en el que se produce el evento terminal*. Y esto, con el objetivo de pronosticar la *probabilidad de que el evento suceda en un momento*

¹ Para profundizar en los contenidos de este capítulo puede consultarse Lee (1992) o Parmar y Machin (1995).

² Especialmente aprovechable es la información de los casos censurados por la *derecha*. No es fácil tratar los casos censurados por la *izquierda* (aquellos de los que se desconoce el momento en que se inicia el seguimiento). En este capítulo se asume que se conoce el momento en el que se inicia el seguimiento de cada caso o que la historia previa del estado de cada sujeto es irrelevante para los objetivos del estudio (para más información sobre tipos de casos censurados y el tratamiento que se les puede dar, ver Cox y Oakes, 1984).

dado del tiempo. En concreto, se intenta pronosticar cuál es la probabilidad de observar un cambio de estado en un momento dado. Por tanto, la *variable objetivo* en un análisis de supervivencia no es si se produce o no el evento, sino el tiempo transcurrido hasta la aparición del evento; es decir, el tiempo de espera o supervivencia.

Existen diferentes formas de abordar el análisis de los tiempos de espera. Aquí revisaremos los tres procedimientos que incluye el SPSS: las **tablas de mortalidad**, el **método de Kaplan-Meier** y el **modelo de regresión de Cox**³. Los dos primeros sirven para lo mismo: obtener curvas de supervivencia y realizar comparaciones entre grupos. Pero difieren en la forma de obtener esas curvas. El método de Kaplan-Meier se basa en los tiempos de espera individuales, las tablas de mortalidad se construyen agrupando los tiempos de espera en intervalos; el primer método es más útil para estimar curvas de supervivencia; el segundo, para obtener los estadísticos que las describen. El modelo de regresión de Cox sirve para pronosticar los tiempos de espera y para identificar las variables que contribuyen a realizar esos pronósticos (igual que un análisis de regresión lineal, pero aprovechando la información que aportan los casos censurados).

Disposición de los datos

Al igual que en un archivo de datos convencional, cada sujeto (cada caso) debe ocupar un registro (una fila) del archivo. En el escenario más simple, que se da cuando el seguimiento de todos los casos comienza al mismo tiempo, es necesario crear dos variables: una, generalmente llamada *estado*, para reflejar el estado en el que se encuentra el sujeto (evento, no-evento) y otra para indicar el momento en el que se ha producido el evento, si es que se produce. Cuando el seguimiento se inicia en momentos distintos es necesario añadir una tercera variable con información sobre el momento en el que se ha iniciado el seguimiento de cada caso (se tienen inicios distintos, por ejemplo, cuando se estudia a pacientes que reciben un determinado tratamiento en momentos distintos).

La variable que informa del estado en el que se encuentra el sujeto toma dos valores. Ya hemos señalado que a estos valores se les suele asignar los códigos 1 y 0 (para el evento y el no-evento, respectivamente).

La variable que informa del momento en que se produce el evento puede ser una variable tipo *fecha*, en cuyo caso indicará el momento en el que se ha producido el cambio de estado o el final del seguimiento, o una variable *numérica*, en cuyo caso indicará el tiempo transcurrido (horas, días, semanas, meses, etc.) desde el comienzo del seguimiento hasta que se produce el cambio de estado o el final del seguimiento.

³ Las tres técnicas son básicamente exploratorias y no paramétricas. Con ellas no se pretende formular un modelo capaz de reproducir exactamente la forma de las funciones sino, más bien, estimar las probabilidades asociadas a los tiempos de espera para llegar a una representación gráfica lo más precisa posible de esas funciones; y esto, sin establecer supuestos acerca de la distribución de los tiempos de espera. Existen aproximaciones paramétricas que se utilizan en áreas como la ingeniería para el estudio de los fallos de producción, el control de calidad, la fatiga de materiales, etc. En estas aproximaciones se intenta encontrar el modelo paramétrico que mejor representa la evolución del evento a lo largo del tiempo. De ahí ha surgido la utilización de distribuciones teóricas como la de Weibull, la exponencial, la de Gompertz, la lognormal, etc. En este capítulo no trataremos estos modelos. Puede encontrarse una buena aproximación a este enfoque en Blossfeld, Hamerle y Mayer (1989), y en Hosmer y Lemeshow (1999).

Supongamos que, en el ejemplo sobre el tiempo que se tarda en abandonar un tratamiento de desintoxicación, la situación real de tres sujetos es la siguiente: los tres sujetos inician el tratamiento el 8 de mayo de 2008. El primer sujeto abandona el tratamiento el 6 de octubre de 2008; el segundo, el 9 de diciembre de 2008; y el tercero sigue con el tratamiento al finalizar el estudio el 17 de febrero de 2009.

En primer lugar, puesto que todos los sujetos comienzan el tratamiento en la misma fecha (08.05.1988), no es necesario registrar el inicio del seguimiento. En segundo lugar, puesto que no todos los sujetos cambian de estado (el tercer sujeto no cambia de estado) y los que cambian no lo hacen en el mismo momento (el sujeto 1 cambia de estado antes que el sujeto 2), el archivo de datos debe construirse con tres registros (uno por cada sujeto) y dos variables: (1) la variable *estado*, con valor 1 para los dos primeros sujetos (los cuales cambian de estado, es decir, abandonan el tratamiento durante el periodo de seguimiento) y valor 0 para el tercer sujeto (que continúa en tratamiento al finalizar el seguimiento; es un caso censurado) y (2) la variable *tiempo*, que recoge el momento exacto en el que se produce el cambio de estado (si se produce) o el final del seguimiento (si es un caso censurado).

La Figura 9.1 muestra, reproducidos en el *Editor de datos* del SPSS, los datos de los tres sujetos del ejemplo. A la variable *estado* le hemos dado formato *numérico* sin decimales; a la variable *tiempo* le hemos dado formato de *fecha (dd.mm.aaaa)*. El archivo contiene una tercera variable llamada *espera* (con formato numérico y sin decimales). El análisis de supervivencia no se basa en fechas como las asignadas a la variable *tiempo*, sino en los *tiempos de espera*. Estos tiempos representan el tiempo transcurrido entre el inicio del seguimiento y el cambio de estado o el final del seguimiento. Para obtener estos tiempos se ha creado la variable *espera* utilizando las fechas de la variable *tiempo* y la fecha de inicio del seguimiento⁴.

Figura 9.1. *Editor de datos* con tres casos

	estado	tiempo	espera
1	1	06-Oct-2008	151
2	1	09-Dec-2008	215
3	0	17-Feb-2009	285

Tablas de mortalidad

Las **tablas de mortalidad**, también llamadas *tablas de vida* y *tablas actuariales*, son el método más antiguo y utilizado para resumir los tiempos de espera. Estas tablas se elaboran a partir de varios estadísticos y funciones que se obtienen combinando los tiempos de espera con la presencia-ausencia del evento estudiado. Para describir este tipo de tablas vamos a servirnos de los datos de la Tabla 9.1. Estos datos corresponden a 100

⁴ Esta nueva variable puede crearse mediante la opción **Calcular** del menú **Transformar** utilizando como expresión numérica: `CTIME.DAYS(TIEMPO - DATE.DMY(08,05,2008))`. Con esta expresión se tienen los tiempos de espera en días.

participantes en un tratamiento de desintoxicación de un año. Los tiempos de espera se han agrupado en meses; la columna *tiempo* indica el mes de observación. El *número de abandonos* es el número de eventos que se producen cada mes. El *número de casos censurados* es el número de sujetos a los que se les ha perdido la pista antes de finalizar el estudio (meses 1 al 11) o que todavía permanecen bajo tratamiento al finalizar el estudio (mes 12).

El correspondiente archivo de datos SPSS tendrá 100 registros (uno por sujeto) y dos variables: *estado* (1 = “evento”, 0 = “censurado”) y *tiempo* (con el tiempo transcurrido hasta el abandono o el final del seguimiento). Estos datos están disponibles en el archivo *Supervivencia abandono tto*, en la página web del manual.

Tabla 9.1. Datos de 100 sujetos sometidos a tratamiento de desintoxicación

<i>Tiempo</i>	<i>Nº abandonos</i>	<i>Nº casos censurados</i>
1	2	1
2	3	0
3	6	2
4	5	1
5	9	0
6	2	1
7	12	2
8	6	1
9	6	0
10	8	3
11	10	1
12	2	17

Para construir una tabla de mortalidad es necesario comenzar dividiendo la variable que define el *tiempo* en k **intervalos**: $I_1, I_2, \dots, I_i, \dots, I_k$ ($i = 1, 2, \dots, k$). Los tiempos de espera de la Tabla 9.1 se han agrupado en 12 intervalos. Estos intervalos no tienen por qué tener la misma amplitud; de hecho, el último intervalo suele ser abierto. Una vez definidos los intervalos, se procede a calcular una serie de estadísticos y funciones especialmente diseñados para describir tiempos de espera:

1. **Número de eventos:** d_i . Número de casos que experimentan el evento (cambian de estado) en cada intervalo de tiempo. En el ejemplo de la Tabla 9.1, el *número de abandonos* que se van produciendo en cada mes.
2. **Número de casos censurados:** c_i . Número de casos a los que se les pierde la pista antes de experimentar el evento (en el ejemplo, los casos censurados de los meses 1 al 11) más el número de casos que en el momento de finalizar el estudio todavía no han experimentado el evento (en el ejemplo, los 17 casos del mes 12). La incorporación de estos casos al análisis es lo que caracteriza al análisis de supervivencia.

3. **Número de sujetos expuestos r_i .** Número de casos que tienen la posibilidad (están en *riesgo*) de experimentar el evento en cada intervalo de tiempo:

$$r_i = n_i - c_i/2 \quad (\text{con } n_1 = n \text{ y } c_0 = 0) \quad [9.1]$$

donde n es el número total de casos que inicia el estudio, n_i es el número de casos que permanecen bajo seguimiento al inicio del intervalo i (casos que no han experimentado el evento ni son casos censurados antes del intervalo i) y c_i es el número de casos censurados en el intervalo i . Para aprovechar la información que pueden aportar a los casos censurados se asume que están homogéneamente distribuidos en el intervalo de observación y que, consecuentemente, han sido observados durante la mitad del intervalo.

En los datos de la Tabla 9.1, el número de casos con riesgo de experimentar el evento en los dos primeros intervalos vale:

$$\begin{aligned} r_1 &= n_1 - c_1/2 = 100 - 1/2 = 99,5 \\ r_2 &= n_1 - c_2/2 = 97 - 0/2 = 97,0 \end{aligned}$$

4. **Proporción de eventos: q_i .** Proporción de casos que experimentan el evento en cada intervalo de tiempo. También se le llama *proporción de casos que terminan*. Se obtiene a partir del número de eventos y del número de casos expuestos:

$$q_i = d_i / r_i \quad [9.2]$$

Las proporciones de eventos de los dos primeros intervalos de la Tabla 9.1 se obtienen de la siguiente manera:

$$\begin{aligned} q_1 &= d_1 / r_1 = 2 / 99,5 = 0,0201 \\ q_2 &= d_2 / r_2 = 3 / 97,0 = 0,0309 \end{aligned}$$

5. **Proporción de no-eventos (p_i).** Proporción de casos que todavía permanecen bajo seguimiento en cada intervalo de tiempo (todavía no han cambiado de estado ni se les ha perdido la pista). Es habitual referirse a esta proporción como *proporción de casos que sobreviven*. Se obtiene a partir del número de casos que cambian de estado en cada intervalo y del número de casos expuestos al inicio de cada intervalo (se trata del valor complementario de la *proporción de eventos*):

$$p_i = 1 - q_i = 1 - d_i / r_i \quad [9.3]$$

En los datos de la Tabla 9.1, las proporciones de no-eventos de los dos primeros intervalos se obtienen de la siguiente manera:

$$\begin{aligned} p_1 &= 1 - d_1 / r_1 = 1 - 2 / 99,5 = 0,9799 \\ p_2 &= 1 - d_2 / r_2 = 1 - 3 / 97,0 = 0,9691 \end{aligned}$$

6. **Proporción acumulada de no-eventos (P_i).** Proporción de casos que siguen bajo seguimiento al final de cada intervalo. Estas proporciones son los tiempos de espera expresados en una escala de 0 a 1:

$$P_i = p_i P_{i-1} \quad (\text{con } P_0 = 1) \quad [9.4]$$

Se utilizan para estimar la curva de supervivencia (ver, más abajo, el párrafo 10). En los datos de la Tabla 9.1, las proporciones acumuladas de no-eventos correspondientes a los dos primeros intervalos valen:

$$\begin{aligned}P_1 &= p_1 P_0 = 0,9799(1) = 0,9799 \\P_2 &= p_2 P_1 = 0,9691(0,9799) = 0,9496\end{aligned}$$

7. **Mediana de los tiempos de espera.** El hecho de que la distribución de los tiempos de supervivencia tienda a ser muy asimétrica (es bastante habitual que unos pocos sujetos tarden mucho más tiempo que el resto en experimentar el evento; o que unos pocos sujetos lo experimenten muy pronto en relación al resto) convierte a la mediana en un estadístico de mayor utilidad que otros promedios.

Ahora bien, si la mediana se calcula de la forma convencional, se obtiene el valor que divide los tiempos de espera en dos mitades (una con el 50 % de los tiempos de espera menores y otra con el 50 % de los tiempos de espera mayores). Y a ese valor se llega sin distinguir entre eventos y casos censurados. Por esta razón la mediana que se utiliza en el análisis de supervivencia no se calcula de la forma convencional. En este contexto la mediana se define como *el valor (tiempo de espera) al que corresponde una proporción acumulada de no-eventos de 0,50*. Puede calcularse de la siguiente manera:

- Si el k -ésimo intervalo (el último intervalo de la serie) deja por encima más de la mitad de los no-eventos, es decir, si $P_k > 0,50$, se considera que la mediana es el límite superior de ese último intervalo: $Mdn = I_{k+1}$.
- Siendo I_i el intervalo en el cual $P_i < 0,50$ y $P_{i-1} \geq 0,50$ (la proporción acumulada de no-eventos es no creciente a lo largo del tiempo), la estimación de la mediana de los tiempos de espera se obtiene mediante

$$Mdn = t_i + \frac{a_{i-1}(P_{i-1} - 0,50)}{P_{i-1} - P_i} \quad [9.5]$$

Aplicando [9.4] a los datos de la Tabla 9.1, se obtiene $P_9 = 0,4586$ (valor menor que 0,50) y $P_8 = 0,5257$ (valor mayor que 0,50). Por tanto, la mediana de los tiempos de espera (el valor que deja por debajo de sí la mitad de los no-eventos) debe encontrarse en el intervalo 9, pues cuando se inicia a ese intervalo todavía *sobreviven* más casos de la mitad (0,5257) y cuando se sale de ese intervalo *sobreviven* menos casos de la mitad (0,4586). Aplicando [9.5] se obtiene

$$Mdn = 9 + \frac{1(0,5257 - 0,5)}{0,5257 - 0,4586} = 9,38$$

Este resultado indica que los sujetos abandonan el tratamiento, en promedio, a los 9,38 meses. La media de estos tiempos de espera vale 7,87. La mediana calculada de la forma convencional vale 8. Y 9,38 es el tiempo de espera que divide en dos partes iguales la distribución de las proporciones de no-eventos: la mitad de los sujetos sobreviven (no abandonan) al menos 9,38 meses.

Además de todos estos estadísticos, al describir los tiempos de espera es habitual recurrir a algunas funciones que aportan información muy útil:

8. **Función de densidad de probabilidad:** $f(t_i)$. Probabilidad de que el evento ocurra entre los momentos t_i y $t_i + h$, para una cantidad h infinitamente pequeña. En términos discretos, probabilidad de que un sujeto cambie de estado en el intervalo i :

$$f(t_i) = P(t = t_i) \quad [9.6]$$

Puede estimarse a partir de la distribución de frecuencias relativas de la variable t , es decir, a partir de la proporción de eventos:

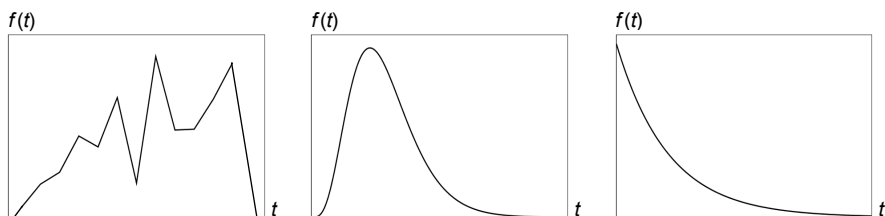
$$\hat{f}(t_i) = (P_{i-1} - P_i) / a_i \quad [9.7]$$

donde a_i se refiere a la amplitud del intervalo i . En los datos de la Tabla 9.1, la densidad de probabilidad de los dos primeros intervalos puede estimarse mediante:

$$\begin{aligned} \hat{f}(t_1) &= (1 - 0,9799) / 1 = 0,0201 \\ \hat{f}(t_2) &= (0,9979 - 0,9496) / 1 = 0,0303 \end{aligned}$$

A la representación gráfica de la función de densidad se le llama *curva de densidad*. La Figura 9.2 muestra varias de estas curvas. La primera de ellas corresponde a los datos de la Tabla 9.1. La curva del centro representa una situación en la que la tasa de eventos es baja al principio, aumenta rápidamente para llegar a su máximo y de nuevo baja rápidamente para tomar valores muy bajos hacia el final. La curva de la derecha representa una situación en la que al principio se produce un tasa muy alta de eventos que va disminuyendo rápidamente conforme va avanzando el tiempo.

Figura 9.2. Ejemplos de funciones de densidad de probabilidad



9. **Función de distribución de probabilidad:** $F(t_i)$. Probabilidad de que el evento ocurra en un momento dado t_i o en cualquier otro anterior él. Se trata, por tanto, de la probabilidad acumulada de eventos hasta el momento i . En el SPSS recibe el nombre de **uno menos la supervivencia**. Puede estimarse a partir de la proporción de casos que han experimentado el evento hasta el intervalo i (incluido ese intervalo) o sumando las probabilidades estimadas para el intervalo i y todos los anteriores a él:

$$\hat{F}(t_i) = P(t \leq t_i) = 1 - P_i = \sum_1^i \hat{f}(t_i) \quad [9.8]$$

En los datos de la Tabla 9.1, la función de distribución correspondiente a los dos primeros intervalos vale (ver los resultados de aplicar [9.7]):

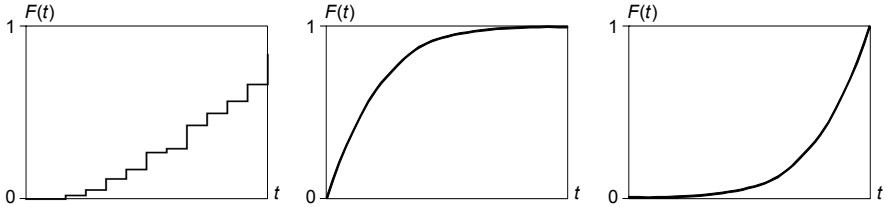
$$\hat{F}(t_1) = \hat{f}(t_1) = 0,0201$$

$$\hat{F}(t_2) = \hat{f}(t_1) + \hat{f}(t_2) = 0,0201 + 0,0303 = 0,0504$$

En el contexto del análisis de supervivencia, una función de distribución puede interpretarse como la probabilidad de que los sujetos *desaparezcan* del seguimiento por haber experimentado un cambio de estado.

La curva de la función de distribución es monótona creciente (ver Figura 9.3) y solo alcanza el valor 1 cuando no existen casos censurados. La primera curva de la Figura 9.3 muestra la función de distribución correspondiente a los datos de la Tabla 9.1 (aparece escalonada porque se basa en intervalos temporales de un mes). La curva del centro refleja una situación en la que los eventos se van produciendo de forma mucho más rápida que en la situación representada en la curva de la derecha. En la última curva la proporción acumulada de eventos empieza a ser alta mucho tiempo después de iniciado el seguimiento.

Figura 9.3. Ejemplos de funciones de distribución de probabilidad



10. **Función de supervivencia: $S(t_i)$.** Función complementaria de la función de distribución de probabilidad:

$$S(t_i) = 1 - F(t_i) \quad [9.9]$$

Se estima a partir de la *proporción acumulada de no-eventos*, es decir, a partir de P_i (ver ecuación [9.4]):

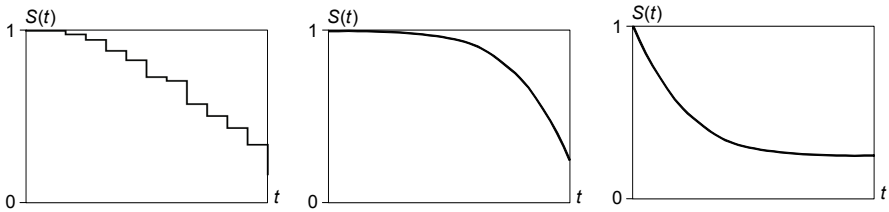
$$\hat{S}(t_i) = P_i = p_i P_{i-1} \quad [9.10]$$

$\hat{S}(t_i)$ es función del tiempo: va disminuyendo conforme avanza el tiempo. Toma su valor máximo, $S(t_i) = 1$, al inicio del seguimiento y su valor mínimo, $\hat{S}(t_i) = 0$, al final, si bien la presencia habitual de casos censurados le impide llegar a 0. Puede interpretarse como la probabilidad de que un sujeto *sobreviva* hasta un momento dado t_i , es decir, como la probabilidad de que el evento no se manifieste hasta el momento t_i .

A la representación gráfica de la función de supervivencia se le suele llamar *curva de supervivencia* y tiene forma monótona decreciente. La Figura 9.4 muestra algunas curvas de supervivencia típicas. La pendiente de la curva indica la intensi-

dad con la que se va produciendo el evento a lo largo del tiempo: a mayor pendiente, mayor intensidad. La curva de la izquierda corresponde a los datos de la Tabla 9.1. La curva del centro representa tiempos de supervivencia muy largos (los eventos se van produciendo lentamente). La curva de la derecha representa tiempos de supervivencia muy cortos (los eventos se producen rápidamente). No es infrecuente encontrar la función de supervivencia representada en escala logarítmica.

Figura 9.4. Ejemplos de funciones de supervivencia



11. **Función de impacto:** $h(t_i)$. Probabilidad condicional de que ocurra el evento en el momento t_i dado que no ha ocurrido antes de ese momento:

$$h(t_i) = \lim_{\Delta \rightarrow 0} [P(t_i \leq t \leq t_i + \Delta t_i \mid t = t_i) / \Delta t_i] \quad [9.11]$$

Recibe diferentes nombres: función de riesgo, tasa de impacto (*hazard rate*), tasa condicional de fallos, tasa de fallos instantánea, tasa de mortalidad condicional, intensidad del fenómeno, etc. Se trata de una medida del riesgo con el que van apareciendo cambios de estado a medida que va avanzando el tiempo. Por tanto, refleja la expectativa de que un caso experimente el evento en un determinado momento. Puede estimarse mediante:

$$\hat{h}(t_i) = \frac{2q_i}{a_i(1 + p_i)} \quad [9.12]$$

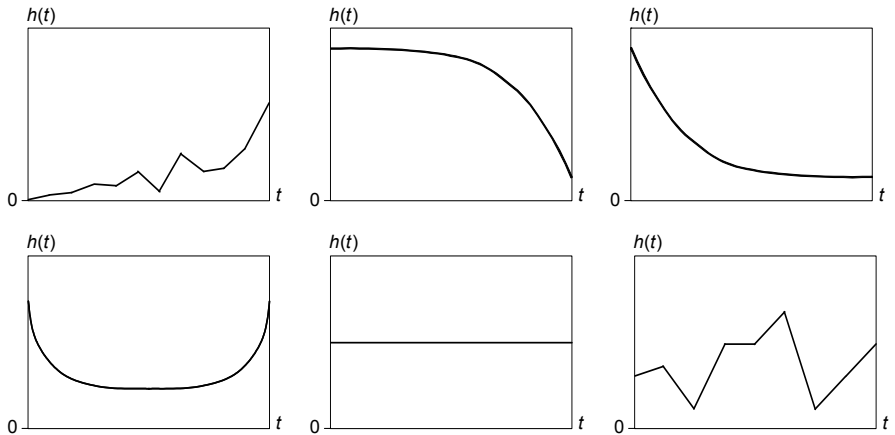
donde q_i es la proporción de eventos en el intervalo i , p_i es la proporción de no-eventos en ese mismo intervalo y a_i es la amplitud del intervalo. Las tasas de impacto correspondientes a los dos primeros meses de la Tabla 9.1 valen:

$$\begin{aligned} \hat{h}(t_1) &= 2q_1/[a_1(1 + p_1)] = 2(2/99,5)/[1(1 + 0,9799)] = 0,0203 \\ \hat{h}(t_2) &= 2q_2/[a_2(1 + p_2)] = 2(3/97,0)/[1(1 + 0,9691)] = 0,0314 \end{aligned}$$

La curva de la función de impacto puede adoptar cualquier forma: creciente, decreciente, constante, etc. La Figura 9.5 muestra distintas funciones de impacto. La primera de ellas corresponde a los datos de la Tabla 9.1. En la segunda curva, la tasa de eventos es muy alta al principio y va disminuyendo con el paso del tiempo (esto es lo que ocurre, por ejemplo, con muchos tratamientos médicos: al principio responden la mayoría de los pacientes y conforme pasa el tiempo van respondiendo los restantes). En la tercera curva, la tasa de eventos también es muy alta al principio,

pero disminuye rápidamente (esto es lo que ocurre, por ejemplo, con el tratamiento de algunos tipos de cáncer: al principio no se responde al tratamiento y la tasa de respuesta va aumentando con el tiempo). La cuarta curva muestra una función decreciente al principio, estable en el centro y creciente al final (esto es lo que ocurre con la tasa de mortalidad de los humanos: las muertes son más numerosas al principio y al final, es decir, entre recién nacidos y ancianos). La siguiente curva representa una función de impacto constante: la tasa de eventos es la misma a lo largo del tiempo (esto es lo que ocurre con la tasa de mortalidad entre los 20 y los 40 años, donde la mayor parte de las muertes se producen por accidente). La última curva refleja una tasa de impacto variable. No es raro que la curva de la función de impacto sea variable, con diversos picos a lo largo del tiempo. Y tampoco es raro que aumente hacia el final del periodo de seguimiento ya que el número de sujetos que permanecen bajo seguimiento puede llegar a reducirse sensiblemente en los momentos finales.

Figura 9.5. Ejemplos de funciones de impacto



Las cuatro funciones expuestas (probabilidad, probabilidad acumulada, supervivencia e impacto) están estrechamente relacionadas. Conociendo una de ellas es posible obtener el resto. La función de supervivencia, por ejemplo, es complementaria de la función de distribución: $S(t_i) = 1 - F(t_i)$. Y la función de impacto se obtiene a partir de las de densidad y supervivencia: $h(t_i) = f(t_i) / S(t_i)$. La función de impacto es, junto con la de supervivencia, la más utilizada e informada en los estudios de supervivencia.

Tablas de mortalidad con SPSS

El procedimiento **Tablas de mortalidad** del SPSS ofrece: (1) las frecuencias, proporciones y funciones necesarias para describir e interpretar los tiempos de espera, (2) representaciones gráficas de esas funciones (densidad, distribución, supervivencia e impacto)

y (3) pruebas de significación para comparar tiempos de espera de distintos grupos⁵. En este apartado se explica cómo obtener toda esta información.

Todos los ejemplos se basan en el archivo *Supervivencia cáncer de mama* (puede descargarse de la página web del manual). El archivo contiene información sobre 1.207 pacientes de cáncer de mama sometidas a tratamiento. Esta información procede de un estudio que intenta describir el tiempo de supervivencia de las pacientes tras ser intervenidas quirúrgicamente. La variable *tiempo* recoge el número de meses transcurridos desde el momento de la intervención hasta el de la muerte, o hasta la finalización del seguimiento o hasta la pérdida de seguimiento. La variable *estado* indica si la paciente ha experimentado el evento (1 = “defunción”) o no (0 = “censurado”). El archivo contiene información adicional que no utilizaremos de momento.

Al analizar tiempos de espera es importante tener en cuenta que la distribución de los tiempos de espera no es una estimación de la función de densidad. Esto solo es así cuando todos los sujetos experimentan el evento, es decir, cuando no existen casos censurados. La Tabla 9.2 muestra la distribución de frecuencias de la variable *estado*: el evento (la muerte) lo ha experimentado el 6% de las pacientes. Si se desea representar la distribución de los tiempos de espera con un histograma, se debe filtrar el archivo seleccionando únicamente los casos que han experimentado el evento; solo de esta manera se obtiene una estimación de la función de densidad. Sin embargo, esta estrategia desaprovecha la información del 94% de los casos del archivo.

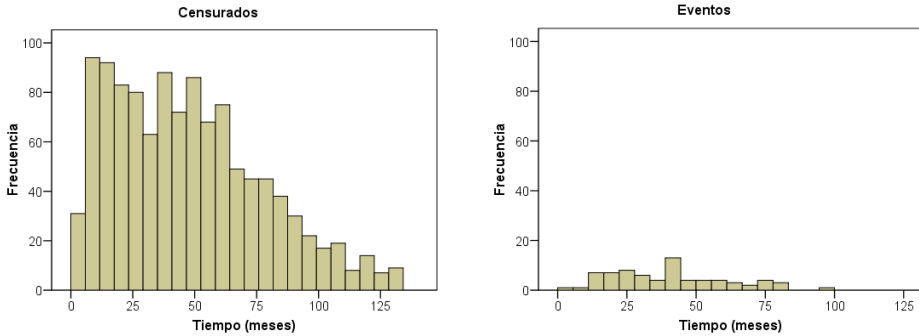
Tabla 9.2. Distribución de frecuencias de la variable *estado*

		Frecuencia	Porcentaje	% válido	% acumulado
Válidos	Censurado	1135	94,0	94,0	94,0
	Muerte	72	6,0	6,0	100,0
	Total	1207	100,0	100,0	

La Figura 9.6 muestra dos histogramas de la variable *tiempo*. El de la izquierda incluye solamente los casos censurados; el de la derecha, solamente los que han experimentado el evento. La forma de ambos histogramas es típica de la distribución de los tiempos de espera: son distribuciones con asimetría positiva. El principal inconveniente de estos histogramas es que no permiten aprovechar de forma conjunta la información de los casos que experimentan el evento y la de los casos censurados. Y ésta es precisamente la ventaja de las tablas de mortalidad: tienen en cuenta los tiempos de espera de todos los casos, hayan experimentado o no el evento. Y lo hacen aprovechando en cada intervalo de tiempo la información de todos los sujetos que llegan a él.

⁵ En una tabla de mortalidad no se establecen supuestos sobre la forma de las funciones que se estiman, pero sí sobre otros aspectos del análisis. En primer lugar, se considera que las probabilidades asociadas al evento sólo dependen del tiempo; por tanto, el momento en que se inicia el seguimiento de cada sujeto no es un aspecto relevante; es decir, se asume que los sujetos que se incorporan al estudio en momentos diferentes (pacientes que inician el tratamiento en momentos distintos, empleados que se incorporan a la empresa en momentos distintos, etc.) se comportan de forma similar. Por otro lado, se asume que los casos censurados y los no censurados no difieren de forma sistemática en ningún aspecto relevante; si el estado clínico o la capacitación laboral, etc., de los casos censurados difiere sistemáticamente del de los no censurados, los resultados estarán, muy probablemente, sesgados.

Figura 9.6. Histogramas de los *tiempos de supervivencia*: casos *censurados* (izqda) y *eventos* (dcha)



Para obtener las frecuencias, proporciones y funciones de una *tabla de mortalidad*:

- Seleccionar la opción **Supervivencia > Tablas de mortalidad** del menú **Analizar** para acceder al cuadro de diálogo *Tablas de mortalidad*.
- Trasladar la variable *tiempo* (tiempo en meses) al cuadro **Tiempo** e introducir el valor 144 en el cuadro **De 0 a** y el valor 12 en el cuadro **por**. Puesto que los tiempos de espera están expresados en meses (valores comprendidos entre 2,63 y 133,80), los valores 144 y 12 permiten abarcar todos los casos y definir intervalos temporales de un año⁶. Puede utilizarse cualquier valor como medida de los tiempos de espera (años, meses, días, etc.). Los casos con valor negativo se excluyen del análisis.
- Trasladar la variable *estado* al cuadro **Variable de estado** (la variable que contiene la información sobre si se ha producido o no el evento; generalmente se tratará de una variable codificada con unos para los eventos y ceros para los casos censurados).
- Para indicar qué código(s) identifica(n) la presencia del evento, pulsar el botón **Definir evento** para acceder al subcuadro de diálogo *Definir evento para la variable de estado* e introducir el valor 1 en el cuadro **Valor único**. La opción **Valor único** es apropiada cuando, como es habitual, el evento está identificado por un solo código (generalmente: 1 = “evento”, 0 = “censurado”). La opción **Rango de valores** es útil cuando el evento está identificado por más de un código; en este caso, los códigos deben ser consecutivos y en los cuadros de texto hay que introducir los límites inferior y superior de ese rango.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestra la Tabla 9.3 (se ha colocado en la cabecera de las columnas la notación utilizada en los apartados ante-

⁶ Para obtener tablas de mortalidad es imprescindible agrupar los tiempos de espera en intervalos. El SPSS utiliza el valor cero como límite inferior del primer intervalo. Los cuadros de texto **De 0 a** y **por** permiten definir el número y amplitud de los intervalos: en el cuadro **De 0 a** es necesario introducir el valor del tiempo de espera más alto que se desea utilizar (lo normal es utilizar el tiempo de espera correspondiente al caso con mayor tiempo de espera); en el cuadro **por** es necesario introducir la amplitud del intervalo. Por ejemplo, si los tiempos de espera se han registrado en meses y el periodo de seguimiento ha durado 4 años, en el archivo de datos se tendrán tiempos de espera comprendidos entre 0 y 48 meses; los valores **e 0 a 48** y **por 6** permitirán crear $48/6 = 8$ intervalos de 6 meses de amplitud.

riores). La primera columna contiene los intervalos de tiempo definidos. Cada intervalo está representado por su límite inferior: I_i = “momento de inicio del intervalo”. La primera fila de datos (intervalo “0”) contiene los casos cuyos tiempos de espera se encuentran entre 0 y 12 meses (los casos cuyo tiempo de espera es exactamente de 12 meses están en la segunda fila). La segunda fila (el intervalo 12) contiene los casos cuyos tiempos de espera están comprendidos entre 12 y 24 meses (los casos cuyo tiempo de espera es exactamente de 24 meses están en la tercera fila). Etc.

La segunda columna (*número que entra en el intervalo*) ofrece el número de casos que continúan bajo seguimiento al inicio de cada intervalo. Al primer intervalo llegan todos los casos incluidos en el análisis: 1.207. Al segundo intervalo llegan 1.076 casos, lo que significa que hay $1.207 - 1.076 = 131$ casos que no continúan en el estudio: 2 de ellos porque han experimentado el evento (*número de eventos*) y 129 porque se les ha perdido la pista y, por tanto, son casos censurados (*número que sale en el intervalo*).

La cuarta columna contiene el número de casos expuestos (*número de expuestos al riesgo*). Se obtiene restando al número de casos que entran en un intervalo (n_i) la mitad del número de casos censurados en ese intervalo ($c_i/2$). Por tanto, en el primer intervalo, $r_1 = 1.207 - (129/2) = 1.142,5$; en el segundo, $r_2 = 1.076 - (183/2) = 984,5$.

Tabla 9.3. Tabla de mortalidad

I_i	n_s	c_i	r_i	d_i	q_i	p_i	P_i	$\hat{f}(t_i)$	$\hat{h}(t_i)$			
Inicio del intervalo ^a	Número que entra en el intervalo	Número que sale en el intervalo	Número expuesto a riesgo	Número de eventos terminales	Proporción que termina	Proporción que sobrevive	Proporción acumulada que sobrevive al final del intervalo	Error típico de la proporción acumulada que sobrevive al final del intervalo	Densidad de probabilidad	Error típico de la densidad de probabilidad	Tasa de impacto	Error típico de tasa de impacto
0	1207	129	1142,5	2	,0018	,9982	,9982	,0012	,0001	,0001	,0001	,0001
12	1076	183	984,5	15	,0152	,9848	,9830	,0041	,0013	,0003	,0013	,0003
24	878	147	804,5	14	,0174	,9826	,9659	,0061	,0014	,0004	,0015	,0004
36	717	166	634,0	20	,0315	,9685	,9355	,0089	,0025	,0006	,0027	,0006
48	531	153	454,5	8	,0176	,9824	,9190	,0105	,0014	,0005	,0015	,0005
60	370	121	309,5	5	,0162	,9838	,9041	,0122	,0012	,0005	,0014	,0006
72	244	91	198,5	7	,0353	,9647	,8723	,0167	,0027	,0010	,0030	,0011
84	146	59	116,5	0	,0000	1,0000	,8723	,0167	,0000	,0000	,0000	,0000
96	87	39	67,5	1	,0148	,9852	,8593	,0209	,0011	,0011	,0012	,0012
108	47	25	34,5	0	,0000	1,0000	,8593	,0209	,0000	,0000	,0000	,0000
120	22	19	12,5	0	,0000	1,0000	,8593	,0209	,0000	,0000	,0000	,0000
132	3	3	1,5	0	,0000	1,0000	,8593	,0209	,0000	,0000	,0000	,0000

a. La mediana del tiempo de supervivencia es 132,00

La proporción de eventos (*proporción que termina*) se obtiene dividiendo el número de eventos del intervalo (d_i) entre el número de sujetos expuestos en ese intervalo (r_i). En el primer intervalo, $q_1 = 2/1.142,5 = 0,0018$; en el segundo, $q_2 = 15/984,5 = 0,0152$.

La proporción de no-eventos (*proporción que sobrevive*) es complementaria de la proporción de eventos. Por tanto, en el primer intervalo, $p_1 = 1 - 0,0018 = 0,9982$; en el segundo, $p_2 = 1 - 0,0152 = 0,9848$.

La columna encabezada *proporción acumulada que sobrevive al final del intervalo* (P_i) ofrece una estimación de la función de supervivencia. Esta estimación se obtiene (ver [9.10]) multiplicando la proporción de no-eventos de cada intervalo, p_i , por la proporción acumulada de no-eventos del intervalo anterior (P_{i-1}). Por tanto, en el primer intervalo, $P_1 = p_1 P_0 = 0,9982(1) = 0,9982$; en el segundo, $P_2 = p_2 P_1 = 0,9848(0,9982) = 0,9830$.

La penúltima columna contiene las estimaciones de la función de impacto: $\hat{h}(t_i)$. Recordemos (ver [9.12]) que esta función se estima mediante: $\hat{h}(t_i) = 2q_i/[a_i(1+p_i)]$. Por tanto,

$$\text{En el intervalo 1: } \hat{h}(t_1) = 2(q_1)/[a_1(1+p_1)] = 2(0,0018)/[12(1+0,9982)] = 0,0001$$

$$\text{En el intervalo 2: } \hat{h}(t_2) = 2(q_2)/[a_2(1+p_2)] = 2(0,0152)/[12(1+0,9848)] = 0,0013$$

Las tres columnas restantes ofrecen los *errores típicos* de las tres principales funciones: supervivencia, probabilidad e impacto. Con muestras grandes, estos errores típicos pueden utilizarse para obtener los intervalos de confianza de los valores individuales de las correspondientes funciones.

Los datos del ejemplo muestran que la proporción acumulada de no-eventos permanece constante en 0,8593 a partir del intervalo 96. Esto está indicando que, pasado ese intervalo, no se produce ningún evento y, consiguientemente, que todos los tiempos de espera posteriores a ese intervalo corresponden a casos censurados.

La proporción de eventos (*proporción que termina*) y la tasa de impacto están relacionadas. La proporción de eventos es una estimación de la función de impacto al final del intervalo, mientras que la tasa de impacto es una estimación de la función de impacto por unidad de tiempo (o estimación promedio dentro del intervalo). Ambos valores siguen patrones similares. Por ejemplo, en el tercer intervalo aparecen 14 eventos. La proporción de eventos al finalizar ese intervalo vale $14/804,5 = 0,0174$, lo cual indica que un caso que sobrevive más allá del segundo año tiene un riesgo del 1,74% de experimentar el evento en el tercer año. Y la tasa de impacto durante el tercer año vale $0,0174/12 = 0,0015$, lo cual indica que el riesgo de experimentar el evento durante un mes cualquiera del tercer año es del 1,5‰.

Una nota a pie de tabla ofrece la *mediana* de los tiempos de espera (132,00). Cuando la mediana toma un valor mayor que el mayor tiempo de espera, se le asigna el valor del último intervalo. Esto es lo que ocurre en el ejemplo. Al finalizar el estudio, al menos el 85,93% de los casos no ha experimentado el evento (ver última fila de la columna P_i); esto significa que no se ha alcanzado el valor de la mediana, el cual corresponde al momento en el que la función de supervivencia toma el valor 0,50; y por esta razón se le asigna el valor 132.

Cómo comparar tiempos de espera

El procedimiento **Tablas de mortalidad** ofrece la posibilidad de comparar los tiempos de espera de varios grupos (por ejemplo, pacientes sometidos a distintos tratamientos, empleados que trabajan en distintas condiciones laborales, etc.) mediante el estadístico de **Wilcoxon-Gehan** (Gehan, 1965a, 1965b; ver Apéndice 9). Este estadístico permite contrastar la hipótesis nula de que las distribuciones de los tiempos de espera de dos o más grupos son iguales.

Siguiendo con el archivo *Supervivencia cáncer de mama*, vamos a comparar los tiempos de espera de los grupos definidos por la variable *tumorcat* (tamaño del tumor, categórica). Esta variable define tres grupos que se diferencian por el tamaño del tumor: 1 = “hasta 2 cm”, 2 = “entre 2 y 5 cm” y 3 = “más de 5 cm”. El primer grupo tiene 826 pacientes, el segundo 238 y el tercero 12. Aunque el tercer grupo solo tiene 12 pacientes, lo vamos a incluir en el análisis para poder ilustrar el proceso de comparación por pares. Para comparar los tiempos de espera de los tres grupos definidos por la variable *tumorcat*:

- Seleccionar la opción **Supervivencia > Tablas de mortalidad** del menú **Analizar** para acceder al cuadro de diálogo *Tablas de mortalidad*.
- Trasladar la variable *tiempo* (tiempo en meses) al cuadro **Tiempo** e introducir el valor 144 en el cuadro **De 0 a** y el valor 12 en el cuadro **por**. Trasladar la variable *estado* al cuadro **Variable de estado**, pulsar el botón **Definir evento** e introducir el valor 1 en el cuadro **Valor único**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Trasladar la variable *tumorcat* (tamaño del tumor) al cuadro **Factor**, pulsar el botón **Definir rango** para acceder al subcuadro de diálogo *Tablas de mortalidad: Definir rango para la variable factor* e introducir los códigos 1 y 3 como valores **Mínimo** y **Máximo**, respectivamente. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Opciones** para acceder al subcuadro de diálogo *Tablas de mortalidad: Opciones* y marcar la opción **Por parejas**⁷ del recuadro **Comparar los niveles del primer factor** y la opción **Supervivencia**⁸ del recuadro **Gráficos**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

⁷ El recuadro **Comparar los niveles del primer factor** contiene las opciones necesarias para efectuar comparaciones entre grupos. La opción **Global** sirve para contrastar la hipótesis nula de que todas las distribuciones poblacionales de los tiempos de espera (tantas como niveles tenga el primer factor seleccionado) son iguales. Si se ha seleccionado un segundo factor, se comparan los niveles del primer factor dentro de cada nivel del segundo. La opción **Por parejas** ofrece comparaciones por pares entre los grupos definidos por los niveles del primer factor (a modo de comparaciones *post hoc*, aunque sin corregir la tasa de error). Si se ha seleccionado un segundo factor, se comparan por pares los niveles del primer factor dentro de cada nivel del segundo.

⁸ Si no se selecciona ningún factor en el cuadro de diálogo principal, el SPSS trata todos los casos del archivo como una única muestra. Si se selecciona un factor, los gráficos incluyen las funciones correspondientes a cada grupo definido por los niveles del factor. Si se selecciona un segundo factor, el SPSS genera un gráfico (con una función para cada subgrupo definido por el primer factor) por cada nivel del segundo factor.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestran las Tablas 9.4 a 9.6 y la Figura 9.7. Además, la tabla de mortalidad (no se muestra aquí) aparece segmentada: una tabla por cada uno de los niveles de la variable *factor*).

La Tabla 9.4 ofrece una comparación *global* de las distribuciones de los tiempos de espera. El estadístico de Wilcoxon-Gehan permite contrastar la hipótesis nula de que las funciones de supervivencia poblacionales de los tres grupos son iguales. El valor del estadístico es 30,02 y tiene asociados 2 grados de libertad (*gl*) y un nivel crítico (*sig.*) menor que 0,0005. Por tanto, se puede rechazar la hipótesis nula y concluir que las funciones de supervivencia comparadas no son iguales.

Tabla 9.4. Comparación global

Estadístico de Wilcoxon (Gehan)	gl	Sig.
30,02	2	,000

La Tabla 9.5 contiene las comparaciones *por pares* entre las tres funciones de supervivencia. La tabla ofrece, para cada una de estas comparaciones, la misma información que la Tabla 9.4 para la comparación global: el estadístico de Wilcoxon-Gehan, sus grados de libertad y su nivel crítico. Los resultados indican que la distribución de los tiempos de espera del grupo 1 difiere significativamente de la del grupo 2 (*sig.* < 0,0005) y de la del grupo 3 (*sig.* = 0,007), y que no existe evidencia de que las distribuciones de los grupos 2 y 3 sean distintas (*sig.* = 0,504).

Tabla 9.5. Comparaciones por pares

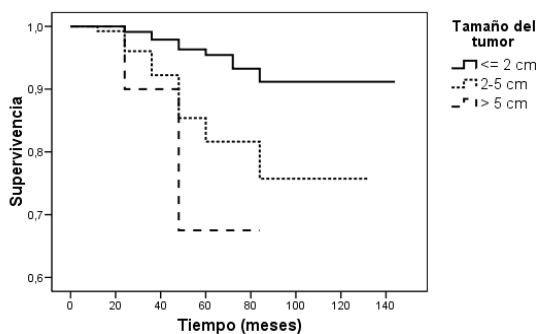
(I) tumorcat	(J) tumorcat	Estadístico de Wilcoxon (Gehan)	gl	Sig.
1	2	27,16	1	,000
	3	7,17	1	,007
2	1	27,16	1	,000
	3	,45	1	,504
3	1	7,17	1	,007
	2	,45	1	,504

La Tabla 9.6 ofrece información descriptiva sobre el tamaño de cada grupo, el número de casos censurados y no censurados, y el porcentaje de casos censurados. También ofrece la *puntuación media* de cada grupo. Para obtener estas puntuaciones medias, el tiempo de espera de cada caso se compara con el de los casos de los restantes grupos; si el tiempo de ese caso es el mayor de los comparados, su puntuación individual aumenta; si es el menor, su puntuación individual disminuye. Las puntuaciones medias de la tabla reflejan el promedio de esas puntuaciones. Y estos promedios indican que los tiempos de espera del primer grupo son mayores, en promedio, que los del segundo, y éstos mayores que los del tercero. Las comparaciones por pares de la Tabla 9.5 ya han permitido concluir que el primer grupo difiere significativamente de los otros dos y que no existe evidencia de que éstos difieran entre sí.

Tabla 9.6. Puntuaciones medias

Grupos	N total	No censurados	Censurados	% de censurados	Puntuación media
1 vs.2 1	826	31	795	96,2%	14,57
2	283	33	250	88,3%	-42,53
1 vs.3 1	826	31	795	96,2%	1,09
3	12	2	10	83,3%	-75,08
2 vs.3 2	283	33	250	88,3%	,47
3	12	2	10	83,3%	-11,08
1	826	31	795	96,2%	1,09
2	283	33	250	88,3%	,47
3	12	2	10	83,3%	-11,08

Finalmente, la Figura 9.7 ofrece en un mismo gráfico curvas de supervivencia separadas para cada uno de los grupos definidos por la variable *tumorcat*. El gráfico permite apreciar con claridad que la curva de supervivencia del grupo 1 = “hasta 2 cm” desciende más lentamente que la de los grupos 2 = “entre 2 y 5 cm” y 3 = “más de 5 cm”. Los resultados del análisis ya han señalado que la curva de supervivencia del grupo 1 difiere de las de los grupos 2 y 3, y que entre las curvas de estos dos grupos no se observan diferencias significativas.

Figura 9.7. Curvas de supervivencia de los tres grupos definidos por la variable *tumorcat*

El método de Kaplan-Meier

El método de *Kaplan-Meier* sirve, al igual que las tablas de mortalidad, para estudiar los tiempos de espera cuando se tienen casos censurados. La característica distintiva de este método es que permite estudiar los tiempos de espera sin necesidad de agruparlos en intervalos, es decir, sin necesidad de establecer cortes de tiempo arbitrarios. En realidad, lo que hace el método de Kaplan-Meier es considerar que los límites de los intervalos son los propios tiempos de espera individuales observados. Por tanto, su lógica es muy parecida a la recién estudiada a propósito de las tablas de mortalidad.

El estadístico *producto-límite*

La Tabla 9.7 resume los datos obtenidos con 10 pacientes enfermos de cáncer sometidos a quimioterapia. La columna *tiempo* contiene los tiempos de espera registrados en semanas. La columna *estado* indica si el tumor ha remitido (1 = “evento”) o no (0 = “caso censurado”).

Tabla 9.7. Datos obtenidos con 10 pacientes de cáncer sometidos a quimioterapia

<i>Tiempo</i>	<i>Estado</i>	$r_i = n_i$	q_i	p_i	$P_i = \hat{S}(t_i)$
9	1	10	0,100	0,900	0,900
12	1	9	0,111	0,889	$0,900 \times 0,889 = 0,800$
13	1	8	0,125	0,875	$0,800 \times 0,875 = 0,700$
18	1	7	0,143	0,857	$0,700 \times 0,857 = 0,600$
18	0	6	0,000	1,000	0,600
23	1	5	0,200	0,800	$0,600 \times 0,800 = 0,480$
28	0	4	0,000	1,000	0,480
31	1	3	0,333	0,667	$0,480 \times 0,667 = 0,320$
45	0	2	0,000	1,000	0,320
122	0	1	0,000	1,000	0,320

Ya sabemos (ver apartados anteriores) que la *proporción de eventos* (q_i) en el momento t_i se obtiene a partir del *número de eventos* (d_i) y del *número de sujetos expuestos* (r_i) en ese momento; es decir: $q_i = d_i/r_i$. Consecuentemente, la *proporción de no-eventos* o de sujetos que *sobreviven* vendrá dada por: $p_i = 1 - d_i/r_i$ (el número de sujetos expuestos, ahora que los datos no están agrupados en intervalos, es simplemente el número de sujetos que permanecen bajo seguimiento en cada tiempo de espera, es decir, n_i). Ahora bien, si los tiempos de espera se registran de forma lo bastante precisa, no existirán empates, en cuyo caso, d_i valdrá 1 para todo i no censurado; por tanto, la proporción de eventos para los casos no censurados podrá calcularse como $q_i = 1/r_i$ y la proporción de no-eventos o de sujetos que *sobreviven* como $p_i = 1 - 1/r_i$. Lógicamente, las proporciones de eventos y no-eventos asociadas a un caso censurado valdrán 0 y 1, respectivamente (ver Tabla 9.7).

Si se asume además que los tiempos de espera están ordenados de forma ascendente (es decir: $t_1 < t_2 < \dots < t_i < \dots < t_n$), la *función de supervivencia* puede estimarse mediante la *proporción acumulada de no-eventos*:

$$\hat{S}(t_i) = P_i = p_i P_{i-1} \quad (\text{con } P_0 = 1) \quad [9.13]$$

Esta forma de estimar la función de supervivencia a partir de las proporciones individuales y acumuladas de no-eventos coincide con el método propuesto por Kaplan y Meier (1958) con su *estimador producto-límite* (ver Lee, 1992, págs. 67-78):

$$\hat{S}(t_i) = \prod_i p_i \quad [9.14]$$

(ver columna $P_i = \hat{S}(t_i)$ en la Tabla 9.7; en el Apéndice 9 se explica cómo construir intervalos de confianza para los valores de la función de supervivencia)⁹.

Cuando el tiempo de espera de un evento coincide con el de un caso censurado, se asume que el evento tiene lugar inmediatamente antes que la censura. Y puesto que la proporción de no-eventos p_i vale 1 cuando no se produce el evento, la función de supervivencia correspondiente a un caso censurado no cambia.

El método de Kaplan-Meier también permite obtener una estimación de la media de los tiempos de espera. Esta estimación refleja el tamaño del área existente bajo la curva de supervivencia y puede calcularse mediante:

$$\hat{\mu} = \begin{cases} \sum_{i=0}^{d-1} \hat{S}(t_i) (t_{i+1} - t_i) & \text{si } t_d = t_n \\ \sum_{i=0}^{d-1} \hat{S}(t_i) (t_{i+1} - t_i) + \hat{S}(t_d) (t_n - t_d) & \text{en otro caso} \end{cases} \quad [9.15]$$

(d se refiere al número de casos no censurados). Por tanto, $d=n$ si no existen casos censurados y $t_d = t_n$ si el tiempo de espera más alto corresponde a un evento. Obsérvese que el sumatorio para obtener $\hat{\mu}$ empieza en el momento 0 (donde $t_0 = 0$ y $\hat{S}(t_i) = 1$) y termina en el penúltimo caso no censurado ($d-1$). Aplicando esta ecuación a los tiempos de espera de la Tabla 9.7 se obtiene:

$$\hat{\mu} = 1(9-0) + 0,900(12-9) + 0,800(13-12) + \cdots + 0,320(122-31) = 51,96$$

Por tanto, el tiempo de supervivencia medio es de, aproximadamente, 52 semanas. Es decir, 52 semanas es el tiempo medio que se estima que los pacientes permanecen bajo tratamiento antes de experimentar la remisión del tumor.

Cuando el tiempo de espera más alto corresponde a un caso censurado, el valor de $\hat{\mu}$ puede estar mal estimado: un caso censurado con un tiempo de espera muy alto podría estar inflando demasiado el valor de la media. En ese caso puede estimarse la media desechando los casos censurados con tiempos de espera mayores que el tiempo de espera correspondiente al último evento (tal como sugiere Irwin, 1949; si se hace esto en el ejemplo se obtiene una media de 22,88). O puede utilizarse la *mediana*. De hecho, el método de Kaplan-Meier permite obtener cualquier cuantil de los tiempos de espera. Siendo p la proporción acumulada de no eventos asociada a un determinado tiempo de espera ($p = 0,25$ para el primer cuartil; $p = 0,5$ para la mediana; etc.):

$$t_p = \inf \left[t_i \mid \hat{S}(t_i) \leq p \right] \quad [9.16]$$

donde *inf* se refiere al tiempo de espera t_i más pequeño para el que la función de supervivencia es igual o menor que p . La mediana, por ejemplo, es el tiempo de espera más pequeño de cuantos acumulan una proporción de no-eventos menor o igual que 0,50. En

⁹ El estimador *producto-límite* puede obtenerse también como un estimador de máxima verosimilitud (ver Kalbfleisch y Prentice, 1980).

el ejemplo de la Tabla 9.7, de todos los valores con $\hat{S}(t_i) < 0,50$, el más pequeño es 23; por tanto, $Mdn = t_{0,50} = 23$. Y el percentil 75 es el tiempo de espera más pequeño de los que acumulan una proporción de no-eventos menor o igual que 0,75; por tanto, $t_{0,75} = 13$.

El método de *Kaplan-Meier* con SPSS

El procedimiento **Kaplan-Meier** permite estimar funciones de supervivencia y calcular la media y la mediana de los tiempos de espera. Para obtener estos estadísticos con los datos de la Tabla 9.7:

- Introducir los datos de la Tabla 9.7 en el *Editor de datos*. Únicamente es necesario introducir las dos primeras columnas; por tanto, solo es necesario crear dos variables; por ejemplo, la variable *tiempo* (con los tiempos de espera) y la variable *estado* (con el estado de cada paciente). Estos datos están disponibles en el archivo *Supervivencia quimioterapia* en la página web del manual.
- Seleccionar la opción **Supervivencia > Kaplan-Meier** del menú **Analizar** para acceder al cuadro de diálogo *Kaplan-Meier* (los cuadros **Tiempo** y **Estado** y el botón **Definir evento** tienen exactamente el mismo significado que en el procedimiento **Tablas de mortalidad**).
- Trasladar la variable *tiempo* al cuadro **Tiempo** y la variable *estado* al cuadro **Estado**.
- Pulsar el botón **Definir evento** para acceder al subcuadro de diálogo *Definir evento para la variable de estado* e introducir el valor 1 en el cuadro de texto correspondiente a la opción **Valor único** (siempre que se hayan respetado los códigos de la Tabla 9.7 al introducir los datos en el SPSS). Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones el *Visor* de resultados ofrece la información que muestran las Tablas 9.8 a 9.10. La primera de ellas (Tabla 9.8) contiene un resumen que incluye el número total de casos incluidos en el análisis, el número de eventos y el número de casos censurados (en frecuencia absoluta y en frecuencia porcentual).

Tabla 9.8. Resumen de los casos procesados

Nº total	Nº de eventos	Censurado	
		Nº	Porcentaje
10	6	4	40,0%

La Tabla 9.9 ofrece la *tabla de supervivencia*; incluye: los tiempos de espera, el estado de cada paciente (evento, censurado), los valores de la función de supervivencia (estimada con estadístico *producto-límite*) o proporción acumulada de no-eventos (no se ofrece para los casos censurados porque, según se ha señalado ya, la función de supervivencia solo cambia entre un evento y el siguiente), los errores típicos de los valores de la función de supervivencia (ver Apéndice 9, al final del capítulo), el número acumu-

lado de eventos en cada momento y el número de sujetos que continúan bajo seguimiento después de cada tiempo de espera.

La Tabla 9.10 incluye información relativa a la media y a la mediana de los tiempos de espera. La media vale 51,96 y aparece acompañada de su error típico¹⁰ y del intervalo de confianza calculado al 95%. Los límites del intervalo de confianza indican que el verdadero tiempo medio que los pacientes permanecen bajo tratamiento antes de experimentar la remisión del tumor se encuentra entre 18,17 y 85,75 semanas. Una nota a pie de tabla recuerda que en el cálculo de la media se tiene en cuenta el mayor tiempo de espera censurado. Puesto que el tiempo de espera más alto correspondiente a un caso censurado (122 semanas) es mucho mayor que el tiempo de espera correspondiente al último evento (31 semanas), el valor de la media es probable que esté muy sesgado. Ya se ha señalado que, para evitar este sesgo, la media podría calcularse desechando los casos censurados ubicados después del último evento. Pero también puede optarse por utilizar la información que proporciona la mediana.

La mitad derecha de la tabla muestra el valor de la mediana acompañado de su error típico y de su intervalo de confianza calculado al 95%. Puede observarse que el valor de la mediana es mucho menor que el de la media. Teniendo en cuenta que el tiempo de espera del último caso (que es un caso censurado) es muy alto en comparación con el del resto de los casos, es realista pensar que el promedio de permanencia bajo tratamiento, antes de remitir el tumor, se parece más a 23 semanas que a 52.

Tabla 9.9. Función de supervivencia

	Tiempo	Estado	Proporción acumulada que sobrevive hasta el momento		Nº de eventos acumulados	Nº de casos que permanecen
			Estimación	Error típico		
1	9,000	Evento	,900	,095	1	9
2	12,000	Evento	,800	,126	2	8
3	13,000	Evento	,700	,145	3	7
4	18,000	Evento	,600	,155	4	6
5	18,000	Censurado	.	.	4	5
6	23,000	Evento	,480	,164	5	4
7	28,000	Censurado	.	.	5	3
8	31,000	Evento	,320	,170	6	2
9	45,000	Censurado	.	.	6	1
10	122,000	Censurado	.	.	6	0

Tabla 9.10. Media y mediana de los tiempos de supervivencia

Media ^a				Mediana			
Estimación	Error típico	Intervalo confianza al 95%		Estimación	Error típico	Intervalo confianza al 95%	
		L. inferior	L. superior			L. inferior	L. superior
51,96	17,24	18,17	85,75	23,00	7,61	8,08	37,92

a. La estimación se limita al mayor tiempo de supervivencia si se ha censurado.

¹⁰ El lector interesado en conocer cómo se calculan estos errores típicos puede consultar Gross y Clark, 1975, o Tarone y Ware, 1977.

Gráficos de los tiempos de espera

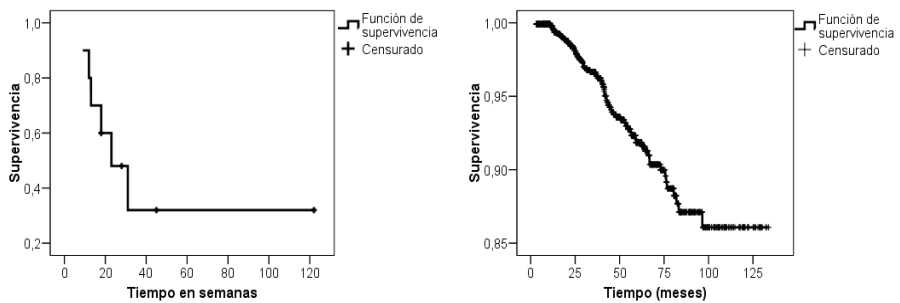
Además de estimar la función de supervivencia y calcular la media y la mediana de los tiempos de espera, el procedimiento **Kaplan-Meier** permite obtener los gráficos típicos de un análisis de supervivencia. Estos gráficos pueden solicitarse en el subcuadro de diálogo *Kaplan-Meier: Opciones*¹¹.

El procedimiento permite obtener cuatro gráficos distintos, todos ellos relacionados con la función de supervivencia. Si no se selecciona ninguna variable *factor* en el cuadro de diálogo principal, el SPSS trata todos los casos del archivo como una única muestra y ofrece una sola curva por gráfico. Si se selecciona una variable *factor*, cada gráfico incluye las curvas correspondientes a los distintos grupos definidos por la variable *factor*. Si además se selecciona una variable *estratos*, el SPSS genera un gráfico por cada nivel de la variable *estratos* y en cada uno de estos gráficos incluye una curva por cada grupo definido por los niveles de la variable *factor*.

Las opciones del recuadro **Gráfico** permiten elegir uno o más de los siguientes gráficos:

1. **Supervivencia** (Figura 9.8). Gráfico con los tiempos de espera en el eje horizontal y los valores de la función de supervivencia (estimados con el estadístico *producto-momento* de Kaplan-Meier) en el eje vertical. La curva que se obtiene es ligeramente distinta de la que se obtiene con el procedimiento **Tablas de mortalidad** por la sencilla razón de que, ahora, los tiempos de espera no están agrupados en intervalos. Pero la lectura del gráfico es similar: la anchura de los escalones representa el tiempo transcurrido entre un evento y el siguiente; la altura refleja cómo va disminuyendo la proporción de no-eventos: con cada evento se produce un descenso. Los casos censurados están representados con cruces. La curva permite hacerse una

Figura 9.8. Funciones (curvas) de supervivencia



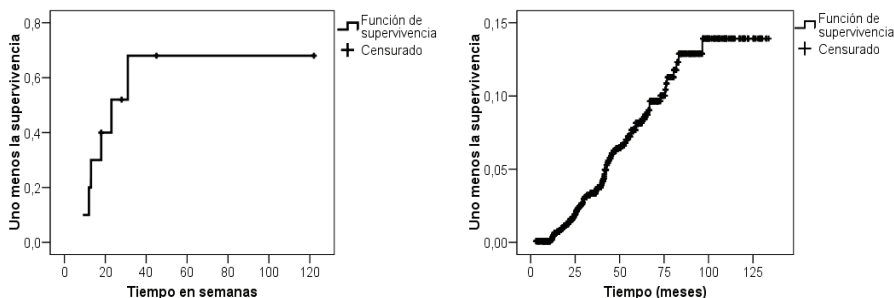
¹¹ Este subcuadro de diálogo también permite decidir qué estadísticos se desea obtener. Todos ellos se ofrecen por defecto, excepto los cuartiles (los percentiles 25, 50 y 75 acompañados de sus respectivos errores típicos). Si se selecciona una variable *factor*, tanto la función de supervivencia como la media, la mediana y los cuartiles se calculan para cada uno de los grupos definidos por los niveles de la variable *factor*. Si además se selecciona una variable *estratos*, tanto la función de supervivencia como la media, la mediana y los cuartiles se calculan para cada uno de los subgrupos resultantes de combinar los niveles de la variable *factor* con los niveles de la variable *estratos*.

idea aproximada de cualquier cuantil de la distribución; la mediana, por ejemplo se encuentra ligeramente por encima de 20 semanas.

La Figura 9.8 muestra dos curvas de supervivencia. La curva de la derecha se ha obtenido con una muestra de casos mucho mayor y con intervalos temporales más pequeños que la de la izquierda. De ahí que en lugar de escalones muestre una mayor continuidad. La forma de la curva informa sobre lo que está ocurriendo: en los primeros 10 meses no se aprecia descenso, indicando esto que en ese periodo de tiempo no se producen eventos; a partir de ese momento se observa un descenso pronunciado y más o menos constante hasta pasados los 80 meses, momento a partir del cual se produce un estancamiento que se prolonga con los casos censurados del tramo final.

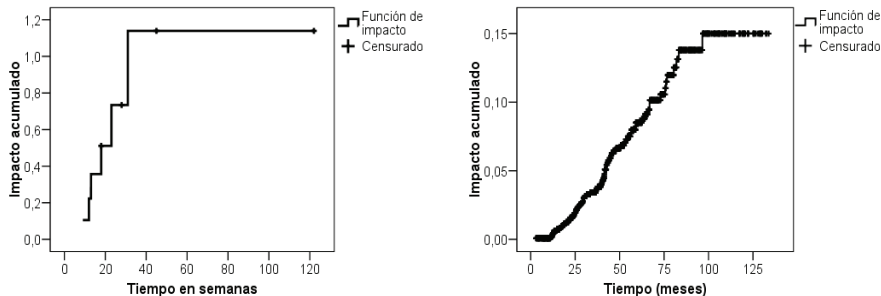
2. **Uno menos la supervivencia** (Figura 9.9). Gráfico de los valores complementarios de la función de supervivencia. Se trata, por tanto, de una representación de la función de distribución $F(t_i)$ o función de probabilidad acumulada. Esta función es igual que la de supervivencia, pero rotada 180° verticalmente.

Figura 9.9. Funciones (curvas) *uno menos la supervivencia*



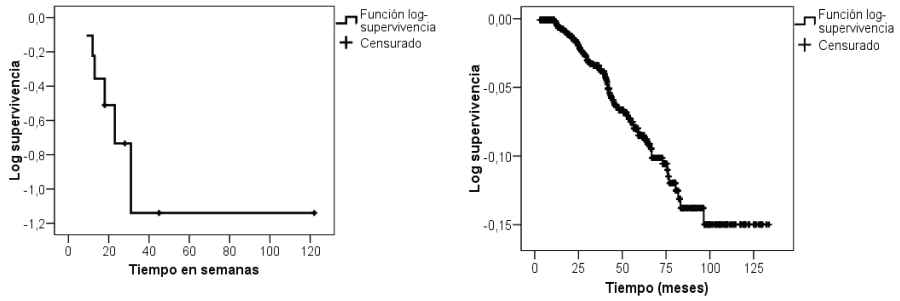
3. **Impacto** (Figura 9.10). Gráfico de la función de impacto acumulado. Refleja la intensidad con la que se van produciendo eventos. Cuanto mayor es la pendiente de esta función, más rápidamente se producen los eventos.

Figura 9.10. Funciones (curvas) de *impacto*



4. **Log de la supervivencia** (Figura 9.11). Gráfico de la función de supervivencia en escala logarítmica. Permite apreciar posibles cambios en el patrón con que los sujetos van desapareciendo del estudio. Una pendiente homogénea indica que los sujetos van desapareciendo de forma constante. Las puntuaciones del eje vertical no tienen valor informativo.

Figura 9.11. Funciones (curvas) *log-supervivencia*



Cómo comparar tiempos de espera

En un análisis de supervivencia es habitual que interese comparar el comportamiento de distintos grupos: pacientes sometidos a distintos tratamientos, empleados que trabajan en distintas condiciones laborales, clientes de distintas áreas geográficas, etc.

El procedimiento **Kaplan-Meier** incluye tres estadísticos para realizar comparaciones entre grupos. Los tres se basan en la diferencia entre el número de eventos observados (n_i) y esperados (m_i) en cada punto temporal. Los tres incluyen un componente general que puede definirse de la siguiente manera:

$$U = \sum_{i=1}^k w_i (n_i - m_i) \quad [9.17]$$

donde k se refiere al número de tiempos de espera distintos y w_i al peso asignado a cada momento i . Los tres estadísticos disponibles en el procedimiento se diferencian en el valor asignado a w_i (ver Lawless, 1982, para una revisión de estos estadísticos). El estadístico **log-rango** (Cox, 1959, 1972; Mantel, 1966; Peto y Peto, 1972) utiliza un peso $w_i = 1$; es decir, todos los eventos reciben la misma ponderación (este estadístico también se conoce como prueba de Mantel-Cox). El estadístico de **Breslow** (Gehan, 1965a, 1965b; Breslow, 1970) utiliza un peso $w_i = r_i$, es decir, pondera cada evento por el número de sujetos expuestos en el momento de producirse el evento; por tanto, los eventos del principio reciben mayor ponderación que los del final, pues el número de sujetos expuestos va disminuyendo conforme pasa el tiempo (este estadístico también se conoce como prueba de Wilcoxon generalizada). Y el estadístico de **Tarone y Ware** (1977) utiliza un peso $w_i = \sqrt{r_i}$, es decir, pondera cada evento por la raíz cuadrada del número de sujetos expuestos en el momento de producirse el evento. Por tanto, los eventos del

principio reciben mayor ponderación que los del final, pero de forma menos acusada que con el estadístico de Breslow. Los tres estadísticos se aproximan a la distribución χ^2 con grados de libertad igual al número de grupos menos 1.

La prueba log-rango es más potente que la de Breslow para detectar diferencias cuando la tasa de mortalidad de un grupo es múltiplo de la del otro grupo (lo que se conoce como tasas de impacto proporcionales; ver, más adelante, en este mismo capítulo, el apartado *Regresión de Cox*). Si no se da esta circunstancia, la prueba de Breslow puede resultar más potente que la prueba log-rango, si bien la de Breslow tiene escasa potencia cuando el porcentaje de casos censurados es muy elevado (Prentice y Marek, 1979). Cuando se realiza un gran número de comparaciones es preferible utilizar del estadístico de Tarone y Ware. Y siempre es recomendable aplicar la corrección de Bonferroni para controlar la tasa de error. En cualquier caso, las distribuciones de los tres estadísticos pueden verse alteradas cuando los patrones de censura de los grupos comparados son muy distintos, especialmente si los tamaños muestrales son pequeños.

Veamos como realizar algunas comparaciones con los datos del archivo *Supervivencia cáncer de mama* (ya lo hemos utilizado para obtener tablas de mortalidad; puede descargarse de la página web del manual). El archivo se ha filtrado utilizando la variable *tumorcat* (tamaño del tumor) para excluir del análisis los casos con un tumor mayor de 5 cm; de este modo, la variable *tumorcat* queda con dos niveles: 1 = “hasta 2 cm” y 2 = “entre 2 y 5 cm”. Para comparar las funciones de supervivencia de estos dos grupos:

- Seleccionar la opción **Supervivencia > Kaplan-Meier** del menú **Analizar** para acceder al cuadro de diálogo *Kaplan-Meier* y trasladar la variable *tiempo* al cuadro **Tiempo**, la variable *estado* al cuadro **Estado** y la variable *tumorcat* (tamaño del tumor) al cuadro **Factor**.
- Pulsar el botón **Definir evento** para acceder al subcuadro de diálogo *Definir evento para la variable de estado* e introducir el valor 1 en el cuadro de texto correspondiente a la opción **Valor único**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Pulsar el botón **Comparar factor**¹² para acceder al subcuadro de diálogo *Kaplan-Meier: Comparar los niveles de los factores* y marcar las opciones correspondientes

¹² El procedimiento **Kaplan-Meier** incluye varias opciones para llevar a cabo distintos tipos de comparaciones entre los niveles de un factor. **Combinada sobre los estratos** contrasta la hipótesis de que todas las funciones de supervivencia poblacionales (tantas como niveles tenga la variable *factor*) son iguales; **Para cada estrato** contrasta la misma hipótesis, pero dentro de cada estrato; **Por parejas sobre los estratos** contrasta la hipótesis de igualdad de funciones de supervivencia comparando por pares los subgrupos definidos por los niveles de la variable *factor* (de modo similar a como se hace con las comparaciones *post hoc* de un ANOVA, aunque sin corregir la tasa de error); **Por parejas en cada estrato** contrasta la hipótesis de igualdad de funciones de supervivencia comparando por pares los subgrupos definidos por los niveles de la variable *factor* dentro de cada estrato (de modo similar a como se hace en las comparaciones *post hoc* de un ANOVA, aunque sin corregir la tasa de error).

Cuando los niveles del factor están cuantitativamente ordenados (dosis de un fármaco, grupos de edad, etc.) y uniformemente espaciados, la opción **Tendencia lineal para los niveles del factor** permite contrastar la hipótesis nula de ausencia de relación lineal entre la función de supervivencia y la variable *factor*. Para contrastar esta hipótesis se utilizan los mismos estadísticos que para realizar el resto de comparaciones. Al marcar esta opción se desactivan las opciones que permiten efectuar comparaciones por pares.

a los tres estadísticos: **Log-rango**, **Breslow** y **Tarone-Ware**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

- Pulsar el botón **Opciones** para acceder al subcuadro de diálogo *Kaplan-Meier: Opciones* y marcar la opción **Supervivencia** del recuadro **Gráficos** (para obtener el gráfico de la función de supervivencia) y desmarcar la opción **Tabla de supervivencia** del recuadro **Estadísticos** (para evitar obtener una tabla de mortalidad demasiado larga y poco informativa). Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones se obtienen los resultados que muestran las Tablas 9.11 a 9.13 y la Figura 9.12. La Tabla 9.11 ofrece información descriptiva que incluye, para cada grupo definido por la variable *factor* y para toda la muestra, el número de casos válidos (*n° total*), el número de eventos y el número de casos censurados (en frecuencia absoluta y porcentual).

Tabla 9.11. Resumen de los casos procesados

Tamaño del tumor	N° total	N° de eventos	Censurado	
			N°	Porcentaje
<= 2 cm	826	31	795	96,2%
2-5 cm	283	33	250	88,3%
Global	1109	64	1045	94,2%

La Tabla 9.12 incluye información sobre las medias de los tiempos de espera acompañadas de sus correspondientes errores típicos e intervalos de confianza. Las medianas no se han podido calcular porque la función de supervivencia (proporción acumulada de no-eventos) no baja hasta el valor 0,50 en el periodo de seguimiento (el número de eventos no alcanza el 50%). El valor de las medias indica que el área existente bajo la curva de supervivencia del grupo “hasta 2 cm” es mayor (126,734) que la del grupo “entre 2 y 5 cm” (108,484).

Tabla 9.12. Medias de los tiempos de supervivencia

Tamaño del tumor	Media ^a			
	Estimación	Error típico	Intervalo de confianza al 95%	
			L. inferior	L. superior
<= 2 cm	126,73	1,28	124,23	129,23
2-5 cm	108,48	3,23	102,15	114,82
Global	123,20	1,31	120,64	125,77

^a. La estimación se limita al mayor tiempo de supervivencia si se ha censurado.

De lo que se trata ahora es de averiguar si la diferencia observada es o no significativa. Para ello, la Tabla 9.13 ofrece los tres estadísticos solicitados: Log-rango, Breslow y Tarone-Ware. Cada estadístico (*chi-cuadrado*) aparece acompañado de sus grados de libertad (*gl*) y de su nivel crítico (*sig.*). Puesto que el valor del nivel crítico es muy pe-

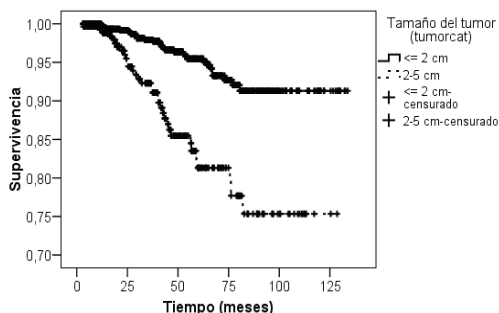
queño (menor que 0,0005 en los tres casos), lo razonable es rechazar la hipótesis nula de igualdad de funciones de supervivencia y concluir que los tiempos de espera de los dos grupos comparados son distintos.

Tabla 9.13. Comparaciones globales (contrastes sobre la igualdad de las funciones de supervivencia)

	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	28,72	1	,000
Breslow (Generalized Wilcoxon)	28,48	1	,000
Tarone-Ware	29,71	1	,000

Las curvas de supervivencia de la Figura 9.12 pueden ayudar a comprender lo que está ocurriendo. En la figura se aprecia claramente que la curva del grupo “hasta 2 cm” desciende más lentamente (por tanto, la supervivencia es mayor) que la curva del grupo “entre 2 y 5 cm”.

Figura 9.12. Curvas de supervivencia



El procedimiento permite utilizar una segunda variable para definir subgrupos dentro de los cuales realizar las comparaciones. Para ello es necesario utilizar una variable que defina *estratos* (seguimos utilizando el archivo *Supervivencia cáncer de mama*):

- En el cuadro de diálogo principal, seleccionar la variable *re* (receptores de los estrógenos) y trasladarla al cuadro **Estratos**.
- Pulsar el botón **Comparar factor** para acceder al subcuadro de diálogo *Kaplan-Meier: Comparar los niveles de los factores* y marcar la opción **Para cada estrato**.

Aceptando estas selecciones se obtienen los resultados que muestran las Tablas 9.14 a 9.16 y la Figura 9.13. La información descriptiva de las Tablas 9.14 y 9.15 se ofrece para cada grupo y estrato: puesto que se está utilizando una variable *factor* con 2 niveles y una variable *estratos* con 2 niveles, los resultados muestran las frecuencias y las medias de los 4 subgrupos resultantes de combinar los dos niveles de ambas variables.

A simple vista, la diferencia entre las medias de los dos grupos definidos por la variable *factor* (*tumorcata*) es ligeramente menor dentro del primer estrato (*negativo*) que

del segundo (*positivo*). En concreto, la diferencia entre los dos grupos en el primer estrato vale $118,878 - 102,728 = 16,15$ y, en el segundo, $127,585 - 108,695 = 18,89$. Las curvas de supervivencia (Figura 9.13) indican esto mismo: la diferencia entre ellas es menos evidente en el primer estrato que en el segundo. La cuestión es si esas diferencias reflejan diferencias poblacionales o son solo fruto del azar muestral. Para ello, la Tabla 9.16 muestra los tres estadísticos solicitados: Log-rango, Breslow y Tarone-Ware. En el primer estrato (*positivo*), los niveles críticos asociados a los tres estadísticos son mayores que 0,05; por tanto, no puede afirmarse que, en ese estrato, las funciones de supervivencia de los dos grupos comparados sean distintas. Sin embargo, en el segundo estrato (*positivo*), los valores de los niveles críticos son, todos ellos, menores que 0,05;

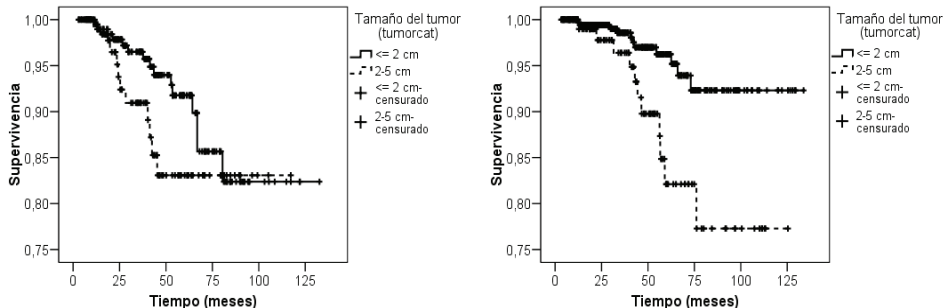
Tabla 9.14. Resumen de los casos procesados

Estado receptores estrógenos	Tamaño del tumor	Nº total	Nº de eventos	Censurado	
				Nº	Porcentaje
Negativo	<= 2 cm	211	15	196	92,9%
	2-5 cm	112	11	101	90,2%
	Global	323	26	297	92,0%
Positivo	<= 2 cm	385	11	374	97,1%
	2-5 cm	119	11	108	90,8%
	Global	504	22	482	95,6%

Tabla 9.15. Medias de los tiempos de supervivencia

Estado receptores estrógenos	Tamaño del tumor	Media			
		Estimación	Error típico	Intervalo de confianza al 95%	
				L. inferior	L. superior
Negativo	<= 2 cm	118,88	3,53	111,95	125,80
	2-5 cm	102,73	4,05	94,79	110,66
	Global	117,81	2,88	112,17	123,44
Positivo	<= 2 cm	127,59	1,93	123,81	131,36
	2-5 cm	108,70	4,67	99,55	117,84
	Global	124,49	2,00	120,57	128,41

Figura 9.13. Curvas de supervivencia. Izquierda: *re = negativo*. Derecha: *re = positivo*.



por lo que se puede rechazar la hipótesis nula de igualdad de funciones de supervivencia y concluir que, en ese estrato, los tiempos de espera de los dos grupos comparados son distintos. La variable utilizada para definir estratos (receptores de los estrógenos) parece ejercer un efecto modulador sobre la aparición de eventos (muertes por cáncer de mama); en concreto: con los receptores de los estrógenos en estado *negativo*, la supervivencia de las pacientes con tumor “hasta 2 cm” no parece que difiera de la de las pacientes con tumor “entre 2 y 5 cm”; por el contrario, con los receptores de los estrógenos en estado *positivo*, las pacientes con tumor “hasta 2 cm” tienen mayor supervivencia que las pacientes con tumor “entre 2 y 5 cm”.

Por supuesto, estos resultados deben interpretarse teniendo en cuenta ciertas consideraciones que ya han sido tratadas en otros capítulos: cuando una diferencia se declara *estadísticamente significativa* se está afirmando que esa diferencia no es atribuible a la pura fluctuación aleatoria del azar muestral; pero esto no significa, necesariamente, que la diferencia encontrada sea *clínicamente relevante*. Las pacientes con tumor “hasta 2 cm” (grupo 1) tienen una supervivencia media entre 16,15 y 18,59 meses mayor que las pacientes con tumor “entre 2 y 5 cm” (grupo 2). Y esto, desde un punto de vista clínico, podría ser mucho más relevante que el hecho de que, dependiendo de que los receptores de los estrógenos se encuentren en estado *positivo* o en estado *negativo*, la diferencia entre las pacientes del grupo 1 y las del grupo 2 sea únicamente de 2,44 meses ($18,59 - 16,15 = 2,44$).

Tabla 9.16. Comparaciones por estrato

Estado receptores estrógenos		Chi-cuadrado	gl	Sig.
Negativo	Log Rank (Mantel-Cox)	1,85	1	,174
	Breslow (Generalized Wilcoxon)	3,34	1	,068
	Tarone-Ware	2,91	1	,088
Positivo	Log Rank (Mantel-Cox)	8,57	1	,003
	Breslow (Generalized Wilcoxon)	6,74	1	,009
	Tarone-Ware	7,98	1	,005

Regresión de Cox

Ya hemos estudiado dos procedimientos alternativos para analizar y representar tiempos de espera: las *tablas de mortalidad* y el *método de Kaplan-Meier*. En este apartado se describe un nuevo procedimiento cuya principal utilidad radica en su capacidad para valorar el efecto de una o más variables independientes o predictoras sobre los tiempos de espera.

El tiempo transcurrido entre dos eventos puede depender de diversos factores. Por ejemplo, el tiempo de recuperación de un paciente diagnosticado de una determinada enfermedad puede depender de la gravedad de la enfermedad, del tipo de tratamiento recibido, de las características del paciente, del tipo de centro hospitalario, etc. Cuando se utiliza una tabla de mortalidad o el método de Kaplan-Meier para valorar el efecto de

una o más variables sobre los tiempos de espera es necesario obtener funciones de supervivencia separadas para cada categoría o combinación de categorías de las variables predictoras. Esta estrategia resulta a menudo insatisfactoria porque el número de casos de cada subgrupo disminuye rápidamente conforme aumenta el número de variables predictoras.

Para resolver este problema podría utilizarse el análisis de regresión lineal con el tiempo de espera como variable dependiente y las variables predictoras como variables independientes. Pero la regresión lineal no tiene en cuenta una característica esencial de los tiempos de espera: la presencia de casos censurados. Cox (1972) ha propuesto una modificación del modelo clásico de regresión lineal que permite analizar los tiempos de espera incorporando la información de los casos censurados.

La ecuación de regresión

La ecuación de regresión de Cox puede formularse de distintas maneras, pero quizá la versión más extendida sea la que utiliza como variable dependiente la función o tasa de impacto: $h(t)$. Esta función ya la hemos definido como la probabilidad condicional de que se verifique un cambio de estado en el momento t , dado que tal cambio no se ha verificado antes de ese momento (ver, en este mismo capítulo, el apartado *Tablas de mortalidad*). Una función de impacto alta indica que la probabilidad de que se produzca el evento es alta (aunque conviene tener presente que una tasa no es una probabilidad; de hecho, una tasa puede tomar valores mayores que uno).

En el caso más simple (una sola variable independiente), el modelo de regresión de Cox adopta la forma:

$$h(t) = h_0(t) e^{\beta X} \quad [9.18]$$

donde X se refiere a la variable independiente (que en el contexto de la regresión de Cox suele recibir el nombre de *covariable*), β es el coeficiente de regresión o peso asignado a la variable independiente, e es la base de los logaritmos naturales y $h_0(t)$ es la función de impacto *basal*, es decir, la función de impacto cuando no se tiene en cuenta la variable independiente, sino solo el tiempo; es decir, cuando $X = 0$. El modelo [9.18] suele expresarse de esta otra manera:

$$h(t)/h_0(t) = e^{\beta X} \quad [9.19]$$

Al cociente $h(t)/h_0(t)$ definido en [9.19] se le suele llamar *impacto relativo* o *razón de impactos* y representa el cambio que se produce en el riesgo de experimentar el evento por efecto de la presencia de la covariable X . Y, al igual que se hace en otros modelos lineales ya estudiados, para facilitar las estimaciones del componente sistemático es habitual tomar logaritmos:

$$\log_e [h(t)/h_0(t)] = \beta X \quad [9.20]$$

Se tiene, de esta manera, un modelo lineal generalizado, con función de enlace logit. El coeficiente β representa el cambio estimado en el logaritmo del *impacto relativo* por cada unidad que aumenta X . Por supuesto, el modelo puede incluir más de una variable independiente o covariable y, al igual que ocurre con el modelo de regresión logística, admite tanto covariables cuantitativas como categóricas. Con p variables independientes adopta la siguiente forma:

$$h(t)/h_0(t) = e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p} \quad [9.21]$$

y tomando logaritmos,

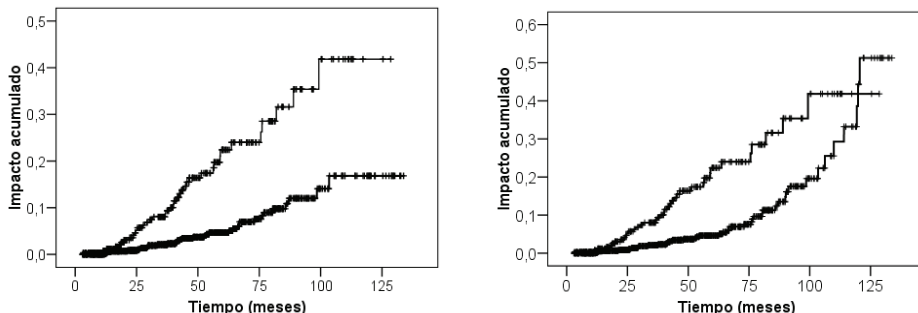
$$\log_e [h(t)/h_0(t)] = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad [9.22]$$

Impacto proporcional

La única diferencia entre las ecuaciones de dos sujetos distintos está en los valores que éstos toman en las covariables. Por tanto, si se dividen las funciones de impacto de dos sujetos, la tasa de impacto basal se anulará, dando lugar a un cociente constante independiente del tiempo (es decir, un cociente que permanecerá constante a lo largo del tiempo). Dicho de otro modo, la ecuación de regresión propuesta por Cox asume que las tasas de impacto de dos sujetos distintos son proporcionales a lo largo del tiempo. De ahí que al modelo de regresión de Cox se le llame *modelo de tasas de impacto proporcionales* o, simplemente, *modelo de impacto proporcional*.

La Figura 9.14 puede ayudar a comprender el concepto de proporcionalidad. En el gráfico de la izquierda están representadas dos funciones de impacto *proporcionales*; aunque la *diferencia* entre ambas funciones no es constante a lo largo del tiempo, a medida que el impacto acumulado va aumentando (a medida que va avanzando el tiempo), el *cociente* entre ambas funciones es aproximadamente el mismo. Cuando dos funciones de impacto son proporcionales, la diferencia entre ellas se va haciendo mayor a medida que van creciendo; por tanto, el supuesto de proporcionalidad entre funciones de impacto implica que sus curvas no se cruzan. Más adelante estudiaremos cómo valorar este supuesto.

Figura 9.14. Funciones de impacto proporcionales (izquierda) y no proporcionales (derecha)



En el gráfico de la derecha están representadas dos funciones de impacto *no proporcionales*. La diferencia entre ellas no va creciendo a lo largo del tiempo. El cociente entre ambas es aproximadamente proporcional hasta la mitad del estudio, momento a partir del cual se pierde la proporcionalidad: la función inicialmente más baja empieza a crecer más rápido que la inicialmente más alta, hasta cruzarla; no existe una razón constante que las relacione. Se tienen covariables dependientes del tiempo, por ejemplo, cuando los pacientes reciben más de un tratamiento durante el periodo de seguimiento; o cuando se toman varias medidas distintas de una misma covariable durante el periodo de seguimiento; etc.

Regresión de Cox con SPSS

El SPSS incluye dos procedimientos para ajustar modelos de regresión de Cox. El procedimiento **Regresión de Cox** permite ajustar modelos de impacto proporcional, es decir, modelos con covariables cuyo efecto se asume constante a lo largo del tiempo. El procedimiento **Regresión de Cox con covariables dependientes del tiempo** permite ajustar modelos con covariables cuyo efecto no se asume constante a lo largo del tiempo (también admite covariables cuyo efecto se asume constante).

Aunque ambos procedimientos utilizan el mismo método de estimación (ambos utilizan el método de máxima verosimilitud y el algoritmo de Newton-Raphson), el primero es preferible al segundo cuando no se tienen covariables dependientes del tiempo, pues incluye varias opciones (algunos gráficos y la posibilidad de guardar algunas variables) que no están disponibles en el segundo.

En este apartado veremos cómo utilizar el SPSS para ajustar modelos de impacto proporcional. En el siguiente apartado veremos cómo ajustar modelos con covariables dependientes del tiempo.

Seguiremos utilizando el archivo *Supervivencia cáncer de mama*, el cual puede descargarse de la página web del manual. Recordemos que el archivo contiene información sobre 1.207 mujeres con cáncer de mama. La variable *estado* indica si se ha producido el evento durante el periodo de tiempo que abarca el estudio (1 = “muerte”, 0 = “censurado”). La variable *tiempo* recoge los tiempos de espera en meses; en las pacientes que han experimentado el evento refleja el tiempo transcurrido entre el inicio del tratamiento y la muerte; en las pacientes que no han experimentado el evento (casos censurados) refleja el tiempo transcurrido entre el inicio del tratamiento y la pérdida del seguimiento o la finalización del estudio. La variable *tiempo* es, junto con la presencia o no del evento, la variable dependiente del análisis, es decir, la variable que se desea pronosticar. Para ajustar un modelo de impacto proporcional:

- Seleccionar la opción **Supervivencia > Regresión de Cox** del menú **Analizar** para acceder al cuadro de diálogo *Regresión de Cox*.
- Trasladar la variable *tiempo* al cuadro **Tiempo** (la variable que contiene los tiempos de espera debe ser *numérica*; puede utilizarse cualquier valor como medida de los tiempos de espera: horas, días, meses, años, etc.).

- Trasladar la variable *estado* al cuadro **Estado** y pulsar el botón **Definir evento** para acceder al subcuadro de diálogo *Regresión de Cox: Definir evento para la variable de estado*; introducir el valor 1 en el cuadro de texto correspondiente a la opción **Valor único** (la variable *estado* puede ser dicotómica o politómica; las opciones de este subcuadro de diálogo permiten indicar qué código(s) de la variable *estado* identifica(n) la presencia del evento).
- Trasladar las variables *edad* (edad en años), *tamaño* (tamaño del tumor en cm), *re* (estado de los receptores de estrógenos), *rp* (estado de los receptores de progesterona) y *lin_sino* (nodos linfáticos positivos) a la lista **Covariables**¹³.

Aceptando estas selecciones se obtienen los resultados que muestran las Tablas 9.17 a 9.21. La Tabla 9.17 contiene información descriptiva. De los 1.207 casos del archivo, solamente 725 han sido incluidos en el análisis. De éstos, en 50 se ha producido el evento y 675 son casos censurados. De los 482 casos excluidos del análisis, la tabla distingue entre los que presentan algún valor perdido en las variables que intervienen en el análisis (392), los que tienen un valor negativo en la variable dependiente *tiempo* (0) y los casos censurados cuyo tiempo de espera es menor que el menor de los tiempos de espera de los casos que experimentan el evento (90). Aunque la tabla no lo indica, unos sencillos estadísticos descriptivos permitirían constatar que la mayor parte de los valores perdidos corresponden a las variables *re* y *rp*; en concreto, de los 392 casos excluidos por tener valor perdido, 338 son casos con valor perdido en *re*, en *rp* o en ambas.

Tabla 9.17. Resumen de los casos procesados

		N	Porcentaje
Casos incluidos en el análisis	Evento	50	4,1%
	Censurado	675	55,9%
	Total	725	60,1%
Casos excluidos	Casos con valores perdidos	392	32,5%
	Casos con tiempo negativo	0	,0%
	Casos censurados antes del primer evento	90	7,5%
	Total	482	39,9%
Total		1207	100,0%

A continuación aparecen los resultados del *Bloque 0*, es decir, los resultados correspondientes al modelo nulo (el modelo que no incluye ninguna covariable). La Tabla 9.18 ofrece el valor del estadístico de ajuste global $-2 \log$ de la verosimilitud (es decir, el estadístico al que venimos llamando desviación y que representamos mediante $-2LL$). El valor de $-2LL$ en el bloque 0 sirve como referente para valorar la contribución al ajuste del conjunto de covariables que se incorporarán al modelo en el siguiente bloque.

¹³ Las covariables pueden tener formato *numérico* o de *cadena corta*; y las variables numéricas pueden ser cuantitativas o categóricas; las variables cuantitativas y las variables dicotómicas pueden introducirse directamente en el análisis, sin embargo, las variables categóricas necesitan un tratamiento especial (ver más adelante el apartado *Variables independientes categóricas*).

Tabla 9.18. Prueba *omnibus* sobre los coeficientes del modelo (Bloque 0)¹⁴

-2 log de la verosimilitud
593,33

La información necesaria para valorar el efecto de las covariables incluidas en el modelo aparece bajo el título *Bloque 1*. La Tabla 9.19 informa del ajuste global del modelo. El hecho de que valor de $-2LL$ haya bajado de 593,33 en el bloque 0 hasta 555,24 en el bloque 1, está indicando que el conjunto de covariables incluidas en el modelo consiguen reducir el desajuste del modelo nulo. Para valorar esa reducción tenemos varios estadísticos *chi-cuadrado*. El estadístico *chi-cuadrado* de la columna *global (puntuación)* contrasta la hipótesis nula de que los coeficientes de todas las covariables incluidas en el modelo valen cero en la población. Puesto que el nivel crítico asociado a este estadístico es muy pequeño (*sig.* < 0,0005), se puede rechazar la hipótesis nula de que todos los coeficientes de regresión valen cero y concluir que existe al menos un coeficiente distinto de cero.

El estadístico *chi-cuadrado* correspondiente al *cambio desde el paso anterior* sirve para valorar el cambio que se produce en el ajuste entre un paso y el siguiente cuando se utiliza un método de ajuste por pasos (ver, más adelante, el apartado *Regresión de Cox por pasos*). Aquí no se ha utilizado un método de ajuste por pasos, de modo que el estadístico *chi-cuadrado* simplemente refleja el cambio experimentado en el ajuste entre el bloque 0 (modelo nulo) y el bloque 1 (modelo propuesto).

Por último, el estadístico *chi-cuadrado* referido al *cambio desde el bloque anterior* refleja la diferencia existente entre las desviaciones de los bloques 0 y 1. El valor de $-2LL$ en el bloque 0 se obtiene fijando en cero todos los coeficientes de regresión (es la desviación del modelo nulo); el valor de $-2LL$ en el bloque 1 se obtiene tras estimar los coeficientes de regresión intentando maximizar la función de verosimilitud parcial. La diferencia entre ambas desviaciones, es decir, la *razón de verosimilitudes*, permite contrastar la hipótesis nula de que las covariables incluidas en el bloque 1 no consiguen reducir el desajuste del modelo nulo. Puesto que esta diferencia (*cambio desde el bloque anterior* = 38,09) tiene asociado un nivel crítico muy pequeño (*sig.* < 0,0005), lo razonable es rechazar la hipótesis nula y concluir que las covariables incluidas en el bloque 1 consiguen reducir significativamente el desajuste del modelo nulo.

Tabla 9.19. Pruebas *omnibus* sobre los coeficientes del modelo (Bloque 1)

-2 log de la verosimilitud ^c	Global (puntuación)			Cambio desde el paso anterior			Cambio desde el bloque anterior		
	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.	Chi-cuadrado	gl	Sig.
555,24	45,93	5	,000	38,09	5	,000	38,09	5	,000

c. Bloque 0: -2 log de la verosimilitud = 593,33

¹⁴ La *verosimilitud parcial* se obtiene calculando el riesgo de aparición del evento en cada momento en el que se produce el evento y multiplicando todos estos riesgos (ver, por ejemplo, Collett, 1994). Este resultado se transforma a escala logarítmica y se multiplica por -2 para obtener el estadístico $-2LL$.

La tabla de *variables incluidas en la ecuación* (ver Tabla 9.20) ofrece las estimaciones de los coeficientes del modelo junto con la información necesaria para valorar su significación y para interpretarlos. Las dos primeras columnas contienen las estimaciones de los coeficientes (*B*) y sus errores típicos (*ET*). Los coeficientes *B* representan el cambio estimado en el *logaritmo del impacto relativo* (variable dependiente) por cada incremento de una unidad en la correspondiente covariable. Los signos positivos de los coeficientes *B* indican que cuando aumenta la covariable, aumenta el impacto relativo; los signos negativos indican que cuando aumenta la covariable, disminuye el impacto relativo. Con las estimaciones que ofrece esta tabla se puede construir la ecuación de regresión:

$$\begin{aligned}\log_e [\hat{h}(t)/\hat{h}_0(t)] &= B_1X_1 + B_2X_2 + \dots + B_pX_p \\ &= -0,02(\text{edad}) + 0,44(\text{tamaño}) - 0,14(\text{re}) - 0,36(\text{rp}) + 0,80(\text{lin_sino})\end{aligned}$$

El estadístico de *Wald* permite contrastar la hipótesis nula de que el correspondiente coeficiente de regresión vale cero en la población. Por tanto, sirve para saber qué covariables tienen un peso significativo en la ecuación. Con covariables cuantitativas y dicotómicas como las del ejemplo, el estadístico de Wald se obtiene elevando al cuadrado el cociente entre el valor del coeficiente (*B*) y su error típico (*ET*), y se distribuye según ji-cuadrado con 1 grado de libertad (*gl*). Con covariables categóricas, el estadístico referido al contraste del efecto global de la variable tiene tantos grados de libertad como el número de categorías menos 1; y el referido al efecto individual de cada categoría tiene 1 grado de libertad.

De las cinco covariables incluidas en el modelo, solamente dos de ellas (*tamaño* y *lin_sino*) tienen asociados coeficientes de regresión significativamente distintos de cero (*sig.* < 0,05). El signo positivo del coeficiente asociado a la covariable *tamaño* indica que la tasa de impacto (el riesgo de que se produzca el evento) es tanto mayor cuanto mayor es el tamaño del tumor. El signo positivo del coeficiente asociado a la covariable *lin_sino* indica que el impacto relativo aumenta cuando se pasa de no tener nodos linfáticos positivos (*lin_sino* = 0) a tenerlos (*lin_sino* = 1).

Para estimar en qué medida cambia el impacto relativo al aumentar el valor de cada covariable, la última columna de la tabla muestra los valores exponenciales de los coeficientes de regresión: $\exp(B)$. Estos valores son *odds ratios* y, consecuentemente, indican cómo cambia el riesgo de experimentar el evento (es decir, cómo cambia el impacto relativo) por cada unidad que aumenta la correspondiente covariable. Los valores obtenidos en nuestro ejemplo permiten concluir lo siguiente: (1) por cada centímetro que

Tabla 9.20. Variables incluidas en la ecuación

	B	ET	Wald	gl	Sig.	Exp(B)
edad	-,02	,01	3,27	1	,070	,98
tamaño	,44	,12	14,09	1	,000	1,55
re	-,14	,38	,14	1	,709	,87
rp	-,36	,38	,92	1	,337	,69
lin_sino	,80	,29	7,50	1	,006	2,22

aumenta el *tamaño* del tumor, el impacto relativo (el riesgo de experimentar el evento) aumenta un 55%; y (2) cuando la covariable *lin_sino* pasa de 0 a 1 (de *no* a *sí*), el impacto relativo queda multiplicado por 2,22 (es decir, aumenta un 122%).

Finalmente, la Tabla 9.21 muestra las medias de las covariables. La media de las covariables dicotómicas codificadas con unos y ceros indica la proporción de unos. En el ejemplo, la proporción de pacientes con nodos linfáticos positivos es de 0,252.

Tabla 9.21. Media de la covariable

	Media
edad	56,04
tamaño	1,82
re	,60
rp	,54
lin_sino	,25

Variables independientes categóricas

Las variables dicotómicas pueden utilizarse como covariables de una regresión de Cox sin necesidad de ninguna consideración adicional; de hecho, en el modelo de regresión del ejemplo anterior ya hemos incluido varias (*lin_sino*, *re*, *rp*). Sin embargo, si una variable categórica tiene más de dos categorías, para poder incluirla como covariable en un modelo de regresión es necesario definirla como categórica y elegir el tratamiento que se le desea dar (los detalles relacionados con este tipo de transformaciones ya se han explicado en el apartado *Covariables categóricas* del Capítulo 5).

Veamos cómo incluir covariables categóricas en un modelo de regresión de Cox. Seguimos utilizando el archivo *Supervivencia cáncer de mama*. Y, como covariable, utilizaremos *tumorcat* (tamaño del tumor categórica), que es una variable categórica con 3 niveles: 1 = “hasta 2 cm”, 2 = “entre 2 y 5 cm” y 3 = “más de 5 cm”.

- En el cuadro de diálogo principal, trasladar la variable *tiempo* al cuadro **Tiempo** y la variable *estado* al cuadro **Estado**.
- Pulsar el botón **Definir evento** para acceder al subcuadro de diálogo *Regresión de Cox: Definir evento para la variable de estado* e introducir el valor 1 en el cuadro de texto correspondiente a la opción **Valor único**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Trasladar la variable *tumorcat* (tamaño del tumor) a la lista **Covariables** y pulsar el botón **Categórica** para acceder al subcuadro de diálogo *Regresión de Cox: Definir covariables categóricas*; trasladar la variable *tumorcat* a la lista **Covariables categóricas**. Dejar **Indicador** como opción del recuadro **Contraste**, seleccionar **Primera** como **Categoría de referencia** y pulsar el botón **Cambiar** para validar la elección hecha. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestran las Tablas 9.22 a 9.24. La Tabla 9.22 muestra la codificación asignada a las categorías

de la variable *tumorcat*. La variable se ha descompuesto en 2 variables *indicador*. A todas las categorías, excepto a la primera, se les ha asignado el valor 1 en la columna correspondiente al parámetro que va a representar a esa categoría en las estimaciones del modelo. El resto de valores en la misma fila y columna son ceros. Esta información sirve para saber que la categoría “entre 2 y 5 cm” va a estar representada por el parámetro o coeficiente 1 y la categoría “más de 5 cm” por el parámetro o coeficiente 2.

Tabla 9.22. Códigos tipo *indicador* asignados a las categorías de la variable *tumorcat*

	Frecuencia	(1)	(2)
tumorcat 1 = <= 2 cm	826	0	0
2 = 2-5 cm	283	1	0
3 = > 5 cm	12	0	1

La Tabla 9.23 contiene las estimaciones de los coeficientes del modelo y su significación. La primera fila (*tumorcat*) ofrece un contraste del efecto global de la covariable *tumorcat*. Si el efecto global no fuera significativo, carecería de sentido seguir inspeccionando el resto de la tabla.

Las siguientes líneas muestran las estimaciones de los coeficientes y su significación. La interpretación que debe hacerse de esta información depende del tipo de codificación elegida, es decir, del tipo de *contraste* elegido para la variable categórica. Un coeficiente significativo (*sig.* < 0,05) indica que la correspondiente categoría difiere significativamente de la *categoría de referencia*. En nuestro ejemplo, dado que hemos elegido una codificación tipo *indicador* con la *primera categoría* como *categoría de referencia*, la línea encabezada *tumorcat(1)* informa sobre la diferencia entre la segunda categoría de *tumorcat* y la primera (que es la categoría de referencia), y la línea encabezada *tumorcat(2)* informa sobre la diferencia entre la tercera categoría y la primera. Puesto que ambos coeficientes tienen asociados niveles críticos (*sig.*) menores que 0,05, se puede afirmar que el impacto relativo de las pacientes con tumores “entre 2 y 5 cm” y “más de 5 cm” difiere del impacto relativo de las pacientes con tumores “hasta 2 cm”. Los valores exponenciales de los coeficientes, *exp(B)*, permiten concretar lo siguiente: (1) el riesgo de muerte es 3,52 veces mayor en la categoría “entre 2 y 5 cm” que en la categoría “hasta 2 cm”; y (2) el riesgo de muerte por cáncer de mama es 6,85 veces mayor en la categoría “más de 5 cm” que en la categoría “hasta 2 cm”.

Tabla 9.23. Variables incluidas en la ecuación

	B	ET	Wald	gl	Sig.	Exp(B)
tumorcat			28,39	2	,000	
Nombre de variable tumorcat(1)	1,26	,25	25,27	1	,000	3,52
Nombre de variable tumorcat(2)	1,92	,73	6,92	1	,009	6,85

Por último, la Tabla 9.24 ofrece las *medias* de las covariables. Con una covariable categórica, estas medias reflejan la proporción de casos que pertenece a cada categoría. A la categoría “entre 2 y 5 cm”, representada por *tumorcat(1)*, pertenece el 25,2% de los

casos; a la categoría “más de 5 cm”, representada por *tumorcata(2)*, pertenece el 1,1 % de los casos; y a la categoría “hasta 2 cm”, que no aparece en la tabla porque es la categoría de referencia, pertenece el $100 - 25,2 - 1,1 = 73,7\%$ de los casos.

Tabla 9.24. Medias de las covariables

	Media
tumorcata(1)	,25
tumorcata(2)	,01

Regresión de Cox por pasos

Cuando, como es habitual, se trabaja con más de una covariable, existen varios métodos para construir un modelo de regresión. El método de **introducción forzosa** genera un modelo de regresión con todas las covariables seleccionadas; esta forma de proceder permite establecer el efecto conjunto de todas las covariables, pero suele conducir a modelos que incluyen covariables que no contribuyen al ajuste. Los métodos de **selección por pasos** permiten utilizar criterios estadísticos para incluir en el modelo únicamente covariables significativas.

Para todo lo relativo a la selección por pasos puede consultarse el apartado *Regresión logística jerárquica o por pasos* del Capítulo 5. Todo lo dicho allí sirve aquí: tanto los métodos de selección de variables como los estadísticos que se utilizan para valorar el ajuste del modelo y la significación estadística de los coeficientes de regresión son los mismos.

Diagnósticos del modelo de regresión de Cox

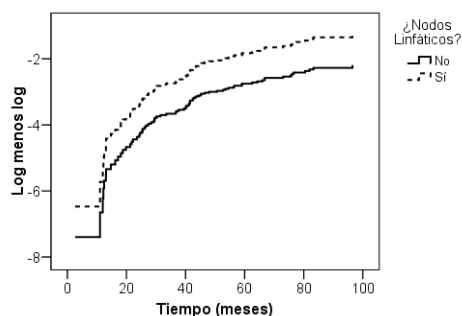
El modelo de regresión de Cox es un modelo de impacto proporcional: se asume que el efecto de las covariables es constante a lo largo del tiempo. Para chequear este supuesto pueden utilizarse diferentes estrategias.

Una muy sencilla consiste en revisar un gráfico llamado *log-menos-log*. El procedimiento **Regresión de Cox** permite obtener varios gráficos: *supervivencia*, *impacto*, *uno menos la supervivencia* y *log-menos-log*. Todos ellos, excepto el último, pueden obtenerse con el procedimiento **Kaplan-Meier** y ya hemos explicado en qué consisten. En el contexto de la regresión de Cox, el gráfico *log-menos-log* tiene una utilidad especial: permite hacer una valoración rápida del supuesto de que las funciones de impacto son proporcionales.

En el archivo *Supervivencia cáncer de mama*, la variable *lin_sino* define dos grupos: el de las pacientes con nodos linfáticos alterados y el de las pacientes con nodos linfáticos no alterados. Si las funciones de impacto de ambos grupos son proporcionales, un gráfico *log-menos-log* (logaritmo del logaritmo negativo de la función de supervivencia) debe mostrar líneas paralelas. En el gráfico de la Figura 9.15 están representadas las funciones *log-menos-log* de los dos grupos definidos por la variable *lin_sino*. El

paralelismo de las líneas indica que no parece haber problemas de proporcionalidad con las funciones de impacto.

Figura 9.15. Función log-menos-log



Casos atípicos e influyentes

Además del gráfico *log-menos-log*, el procedimiento ofrece algunos estadísticos que no solo son útiles para chequear el supuesto de proporcionalidad sino que sirven para valorar la calidad del modelo. Estos estadísticos (residuos, diferencia en las betas) pueden guardarse como variables del archivo de datos mediante las opciones del subcuadro de diálogo *Regresión de Cox: Guardar nuevas variables*.

Residuos de Cox-Snell

Son los valores de la *función de impacto acumulada* (Cox y Snell, 1968). Estos residuos (*haz#* en SPSS) están estrechamente relacionados con los **residuos de martingala**, un tipo particular de residuos tradicionalmente utilizados en el contexto de los modelos de impacto proporcional (ver Therneau, Grambsch y Fleming, 1990). El SPSS no calcula estos residuos, pero el residuo de martingala de un caso censurado no es más que el valor negativo del residuo de Cox-Snell; y el de un caso no censurado se obtiene restando a uno el residuo de Cox-Snell.

La representación de los residuos de martingala respecto de los pronósticos lineales ($X \cdot \text{Beta}$ en el SPSS) ofrece pistas muy valiosas sobre la presencia de casos que el modelo no pronostica bien (probablemente casos atípicos que hay que revisar).

Residuos parciales

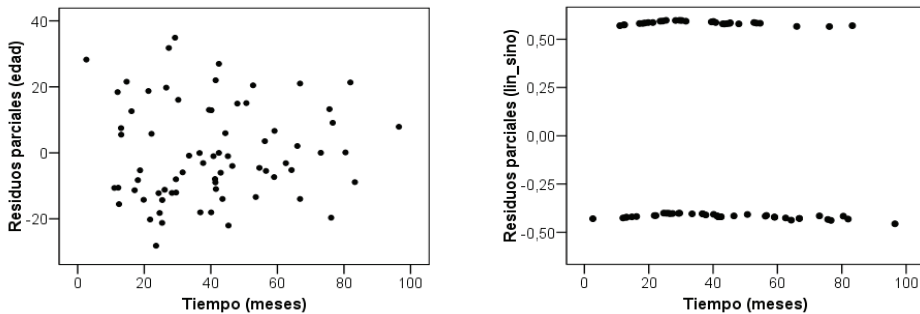
El residuo *parcial* de un caso, también llamado residuo de Schoenfeld (1982; ver también Grambsch y Therneau, 1994), es la diferencia entre su valor observado en una co-variable y su correspondiente valor esperado. Por tanto, a cada caso le corresponden tantos residuos parciales como covariables incluya el modelo de regresión (el SPSS les

asigna los nombres $PR1_#$, $PR2_#$, ..., $PRp_#$). El valor esperado de un caso se calcula a partir del número de casos expuestos cuando el caso en cuestión cambia de estado. Estos residuos no se calculan para los casos censurados, sino solo para los que experimentan el evento.

Los residuos parciales son independientes del tiempo (ver Hess, 1995) y esto hace de ellos una herramienta bastante útil para contrastar el supuesto de proporcionalidad. En condiciones de proporcionalidad, un diagrama de dispersión con el tiempo en el eje de abscisas y los residuos parciales en el de ordenadas debe mostrar una nube de puntos oscilando aleatoriamente en torno a cero.

La Figura 9.16 muestra dos diagramas de dispersión con el *tiempo* en el eje horizontal y los *residuos parciales* en el vertical. En el gráfico de la izquierda están representados los residuos obtenidos con la covariable *edad*; en el de la derecha, los obtenidos con la covariable *lin_sino*. El diagrama de la *edad* muestra residuos aleatoriamente dispersos en torno al valor cero; por tanto, no parece que se esté incumpliendo el supuesto de proporcionalidad. El diagrama de *lin_sino* muestra una nube de puntos organizada en torno a dos filas. Puesto que la covariable *lin_sino* es dicotómica, únicamente puede generar dos posibles pronósticos: los pronósticos basados en $lin_sino = 1$ son positivos y los correspondientes residuos están organizados en torno a 0,60; los pronósticos basados en $lin_sino = 0$ son negativos y los correspondientes residuos están organizados en torno a -0,40. Esta nube de puntos es típica de las covariables dicotómicas que generan tasas de impacto proporcionales; el alejamiento de esta pauta indica falta de proporcionalidad.

Figura 9.16. Diagrama de dispersión con los *residuos parciales*



Diferencia en las betas

La *diferencia en las betas* ($DfBetas$) refleja el cambio que se produce en el tamaño de los coeficientes de regresión al ir eliminando casos uno a uno (el SPSS crea tantas variables nuevas como covariables y las nombra $DFB1_#$, $DFB2_#$, ..., $DFBp_#$).

Al igual que en otros modelos de regresión, las diferencias en las betas informan sobre la posible presencia de casos influyentes (casos con demasiado peso en la ecuación de regresión). En regresión lineal, por ejemplo, las diferencias en las betas se tipifican

para poder interpretarlas (valores tipificados muy grandes o muy pequeños delatan posibles casos influyentes). Aquí, las diferencias en las betas no son tan exactas como en regresión lineal y sus valores tipificados pueden ser muy grandes o muy pequeños incluso con casos poco o nada influyentes.

Covariables dependientes del tiempo

La Figura 9.14 muestra funciones de impacto proporcionales (gráfico de la izquierda) y no proporcionales (gráfico de la derecha). En éste, mientras el impacto acumulado de los sujetos no tratados (línea continua) va creciendo progresivamente a lo largo del tiempo, el impacto acumulado de los sujetos tratados (línea discontinua) crece muy despacio hasta la mitad del seguimiento y comienza a crecer mucho más deprisa a partir de ese momento. El cociente entre los impactos de ambos grupos no se mantiene constante a lo largo del tiempo. Las funciones *log menos log* reflejarían esta circunstancia con curvas no paralelas.

Cuando el efecto de una covariable depende del tiempo, los resultados obtenidos con el modelo de regresión de Cox pueden ser radicalmente distintos de los obtenidos con un modelo de regresión que tenga en cuenta el efecto del tiempo. Por tanto, puesto que el modelo de impacto proporcional de Cox asume que el efecto de las covariables sobre el impacto relativo es constante a lo largo del tiempo, no debe utilizarse para valorar el efecto de una covariable que genera tasas de impacto no proporcionales.

Esta idea puede entenderse fácilmente con un sencillo ejemplo. Imaginemos que se desea valorar el efecto de dos tratamientos distintos. Si el efecto de ambos tratamientos es constante a lo largo del tiempo, la diferencia entre ambos podrá ser establecida sin necesidad de tomar el tiempo en consideración. Por el contrario, si el efecto de los tratamientos cambia con el tiempo, la valoración correcta de la diferencia entre ambos no podrá prescindir de algún tipo de referencia al tiempo.

Una covariable cuyo efecto no es constante a lo largo del tiempo puede incorporarse al análisis creando una *covariable dependiente del tiempo*; es decir, combinando el tiempo o alguna función del tiempo con la covariable que genera tasas de impacto no proporcionales. Esto implica incorporar el tiempo (o alguna función del tiempo) como variable independiente:

$$h(t) = h_0(t) e^{B_1 X + B_2 X(T_COV)} \quad [9.23]$$

Desarrollando y tomando logaritmos:

$$\log_e [h(t) / h_0(t)] = B_1 X + B_2 X(T_COV) \quad [9.24]$$

El coeficiente B_2 es el peso asignado a una covariable formada por el producto de la covariable X (que ya forma parte de la ecuación como covariable con el coeficiente B_1) y una variable llamada T_COV (nombre que el SPSS asigna a la covariable dependiente del tiempo). La variable T_COV no es más que el tiempo (variable dependiente del análisis) o alguna función del tiempo.

Cómo crear covariables dependientes del tiempo

El procedimiento **Regresión de Cox con covariables dependientes del tiempo** permite, por un lado, ajustar modelos con covariables cuyo efecto no se asume constante a lo largo del tiempo y, por otro, incluir covariables cuyos valores no son constantes a lo largo del tiempo, es decir, covariables que toman valores distintos en momentos distintos del seguimiento. Esta segunda circunstancia se da, por ejemplo, cuando los pacientes cambian de tratamiento a lo largo del seguimiento (reciben más de un tratamiento), o cuando se toman varias medidas de una misma covariable en momentos distintos del seguimiento. Para crear covariables dependientes del tiempo:

- Seleccionar la opción **Supervivencia > Regresión de Cox con covariables dependientes del tiempo** del menú **Analizar** para acceder al cuadro de diálogo *Calcular covariable dependiente del tiempo*.

Este cuadro de diálogo previo al principal permite definir la covariable dependiente del tiempo, es decir la covariable *T_COV_* (es posible definir e incluir en el análisis más de una covariable dependiente del tiempo mediante sintaxis).

La lista de variables del archivo de datos incluye una variable adicional llamada “*T_*”. Esta variable no pertenece al archivo de datos; la crea el SPSS haciéndole tomar los mismos valores que la variable del archivo de datos que más adelante será definida (en el cuadro de diálogo principal) como variable dependiente del modelo de regresión.

El cuadro **Expresión para T_COV_** puede utilizarse para definir dos tipos de covariables dependientes del tiempo:

1. Cuando una covariable genera **tasas de impacto no proporcionales** puede crearse e incluirse en el análisis una nueva variable que sea función del tiempo y de la covariable en cuestión. Esta estrategia permite, por un lado, ajustar modelos de impacto no proporcional y, por otro, comprobar si efectivamente las tasas de impacto son no proporcionales.

Este tipo de covariables dependientes del tiempo se suelen crear simplemente multiplicando la covariable en cuestión por la variable *T_*. El resultado de esta transformación se almacena en la variable *T_COV_*, la cual aparece listada como una variable más en el cuadro de diálogo principal. De este modo, al incluir en el análisis la covariable *T_COV_*, se está incluyendo una covariable que es el resultado de multiplicar el tiempo por la covariable que genera tasas de impacto no proporcionales.

2. Cuando se tienen **covariables que toman valores distintos a lo largo del seguimiento** (como, por ejemplo, cuando se cambia de tratamiento, o cuando se toman varias medidas de la misma variable y se desea utilizar la más reciente en cada momento del seguimiento), pero que no están sistemáticamente relacionadas con el tiempo, puede crearse una covariable dependiente del tiempo *segmentada*.

Imaginemos que en el ejemplo que se viene utilizando sobre *cáncer de mama*, la valoración del estado de los nodos linfáticos se ha llevado a cabo en tres momentos distintos del seguimiento: *linpos_1*, *linpos_2* y *linpos_3*. Puesto que no todos los

pacientes tienen el mismo tiempo de supervivencia, habrá pacientes con tres medidas (porque han conseguido sobrevivir hasta el momento en el que se ha efectuado la tercera medida), pero habrá quienes tengan solo dos medidas o solo una (porque no han conseguido sobrevivir hasta el momento de tomar la tercera medida o hasta el momento de la segunda). Si se desea utilizar la medida apropiada en cada tiempo de espera puede crearse una covariable dependiente del tiempo *segmentada*.

Para crear esta covariable segmentada pueden utilizarse las expresiones lógicas de la sintaxis SPSS. Supongamos que la primera medida se ha llevado a cabo al comienzo del estudio, la segunda a los 30 meses y la tercera a los 60 meses. En la expresión:

$$(T_ < 30) * \text{limpos_0} + (T_ \geq 30 \text{ and } T_ < 60) * \text{limpos_30} + (T_ \geq 60) * \text{limpos_60}$$

cada paréntesis arroja un valor de 1 o de 0 dependiendo de que el contenido del paréntesis sea verdadero o falso. El primer paréntesis vale 1 cuando la variable $T_$ (que asume los valores de la variable dependiente *tiempo*) es menor que 30 meses y 0 en cualquier otro caso; el segundo paréntesis vale 1 cuando $T_$ es mayor o igual que 30 y menor que 60 y 0 en cualquier otro caso; el tercer paréntesis vale 1 cuando $T_$ es mayor o igual que 60 y 0 en cualquier otro caso. En consecuencia, la variable $T_COV_$ tomará los valores de *limpos_0* cuando $T_$ sea menor que 30, los valores de *limpos_30* cuando $T_$ tome valores entre 30 y 60, y los valores de *limpos_60* cuando $T_$ sea igual o mayor que 60.

Puesto que los casos con valor perdido son excluidos del análisis, debe tenerse en cuenta que todos los casos válidos deben puntuar en las tres medidas tomadas (*linpos_1*, *linpos_2* y *linpos_3*), incluso aunque el tiempo de supervivencia de un caso sea menor que el tiempo en el que se ha tomado alguna de las tres medidas. Un caso en el que se haya producido el evento (o al que se le haya perdido la pista) a los 20 meses solo tendrá valor válido en *linpos_1*; pero, para no quedar fuera del análisis como consecuencia de la aplicación de una expresión lógica como la propuesta más arriba, es necesario asignarle un valor válido tanto en *linpos_2* como en *linpos_3* (no importa el valor concreto que se le asigne, pues ese valor no será utilizado en el análisis; pero debe ser un valor válido). Esto puede hacerse, por ejemplo, en una ventana de sintaxis mediante sentencias del tipo:

```
If missing (linpos2) linpos2 = 0
```

```
If missing (linpos3) linpos3 = 0
```

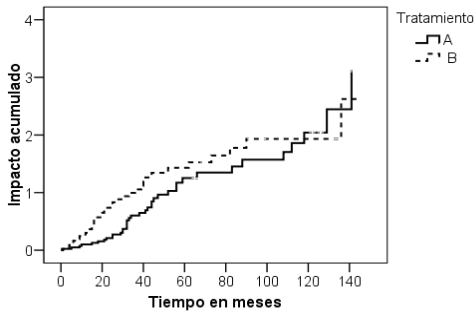
En los siguientes apartados se ofrecen dos ejemplos. En el primero de ellos se utiliza un modelo de tasas de impacto no proporcionales. El segundo ilustra cómo utilizar covariables dependientes del tiempo.

Regresión con covariables dependientes del tiempo

El archivo *Supervivencia infarto* (puede descargarse de la página web del manual) contiene información sobre 88 pacientes con alto riesgo de infarto, 42 de ellos sometidos al tratamiento A y 46 al tratamiento B. El objetivo del análisis es valorar si los tratamien-

tos *A* y *B* difieren en eficacia. Si se pasara por alto el hecho de que los tratamientos *A* y *B* podrían generar tasas de impacto no proporcionales, es decir, si se aplicara a estos datos un modelo de regresión como el utilizado en los ejemplos anteriores, se obtendría un coeficiente de regresión igual a 0,348, con un nivel crítico asociado de 0,134. Este resultado llevaría a concluir que no es posible afirmar que el efecto del tratamiento *A* sea distinto del efecto del tratamiento *B*. No obstante, la función de impacto acumulado de la Figura 9.17 sugiere que el impacto relativo bajo los tratamientos no es independiente del tiempo: el cruce de las líneas indica que las funciones de impacto (una por tratamiento) no son proporcionales.

Figura 9.17. Funciones de impacto acumulado



Así las cosas, valorar apropiadamente el efecto de los tratamientos requiere incorporar al análisis el tiempo como variable independiente. Para hacer esto,

- En el cuadro de diálogo previo al principal seleccionar la variable *T_* y trasladarla al cuadro **Expresión para T_COV_** (de este modo, la covariable *T_COV* tomará los mismos valores que la variable *tiempo*).
- Pulsar el botón **Modelo** para acceder al cuadro de diálogo principal¹⁵; seleccionar la variable *tiempo* y trasladarla al cuadro **Tiempo**; seleccionar la variable *estado* y trasladarla al cuadro **Estado**.
- Pulsar el botón **Definir evento** para acceder al subcuadro de diálogo *Regresión de Cox: Definir evento para la variable de estado* e introducir el valor 1 en el cuadro de texto correspondiente a la opción **Valor único**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- Trasladar la variable *tto* (tratamiento) a la lista **Covariables**. Trasladar las variables *tto* (tratamiento) y *T_COV* a la lista **Covariables** utilizando el botón **>a*b>**.

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestra la Tabla 9.25. Ahora, el coeficiente de regresión estimado para la variable *tto* (tratamiento)

¹⁵ La variable *T_COV_* toma los valores definidos para ella en el cuadro de diálogo previo. El resto del cuadro de diálogo es exactamente igual que el cuadro de diálogo *Regresión de Cox*, con dos excepciones: ahora no está disponible el botón **Gráficos**; y en el botón **Guardar** únicamente está disponible la opción **DfBetas**.

tiene asociado un nivel crítico muy pequeño ($Sig.=0,007$). A diferencia de lo que ocurre cuando no se tiene en cuenta el tiempo, el coeficiente asociado a la covariable *tto* es distinto de cero; lo que significa que los tratamientos difieren en su efecto sobre el impacto relativo. El valor exponencial del coeficiente de regresión, $Exp(B) = 2,81$, indica que la tasa de infartos es casi tres veces mayor entre los pacientes que reciben el tratamiento *B* ($tto = 2$) que entre los que reciben el tratamiento *A* ($tto = 1$).

También el coeficiente de regresión asociado a la interacción entre el tratamiento y el tiempo ($tto * T_COV$) es distinto de cero ($sig.=0,020$), lo cual significa que el efecto de los tratamientos no es independiente del tiempo. Este resultado indica que el modelo que incluye la variable *tto* incumple el supuesto de impacto proporcional y que, consecuentemente, para poder obtener una correcta valoración del efecto de los tratamientos mediante un modelo de regresión de Cox es recomendable recurrir a una covariable dependiente del tiempo.

Tabla 9.25. Variables incluidas en la ecuación

	B	ET	Wald	gl	Sig.	Exp(B)
tto	1,03	,38	7,34	1	,007	2,81
T_COV_*tto	-,02	,01	5,22	1	,022	,98

Regresión con covariables cuyos valores cambian con el tiempo

Este ejemplo muestra cómo utilizar una covariable dependiente del tiempo *segmentada*. El archivo *Supervivencia infarto* utilizado en el ejemplo anterior recoge tres medidas del nivel de colesterol total en sangre (en mg/dl): la primera medida se ha tomado al inicio del seguimiento (*colesterol_0*); la segunda, a las 30 semanas (*colesterol_30*); la tercera, a las 60 semanas (*colesterol_60*). Si se desea utilizar el nivel de colesterol en sangre como variable independiente en un análisis de regresión y se quiere, además, que la medida utilizada con cada sujeto sea la más reciente, es necesario crear una variable *segmentada*. Para ello:

- En el cuadro de diálogo previo al principal, escribir en **Expresión para T_COV_**:
 $(T_ < 30) * coltot_0 + (T_ \geq 30 \text{ and } T_ < 60) * coltot_30 + (T_ \geq 60) * coltot_60$
- Pulsar el botón **Modelo** para acceder al cuadro de diálogo principal y trasladar la variable *tiempo* al cuadro **Tiempo**, la variable *estado* al cuadro **Estado** y la variable *T_COV* a la lista **Covariables**.
- Pulsar el botón **Definir evento** para acceder al subcuadro de diálogo *Regresión de Cox: Definir evento para la variable de estado* e introducir el valor 1 en el cuadro de texto correspondiente a la opción **Valor único**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece, entre otros, los resultados que muestra la Tabla 9.26. El signo positivo del coeficiente (0,015) indica que la relación entre el nivel

de colesterol y la tasa de infartos es positiva. Y, puesto que el correspondiente nivel crítico es muy pequeño ($\text{sig.} < 0,0005$), puede afirmarse que la relación es significativa. El valor exponencial del coeficiente, $e^{10(0,014)} = 1,15$, indica que, por cada 10 mg/dl que aumenta el nivel de colesterol, la tasa de impacto relativo (el riesgo de sufrir infarto) aumenta un 15%.

Tabla 9.26. Variables incluidas en la ecuación

	B	ET	Wald	gl	Sig.	Exp(B)
T_COV_	,014	,003	25,241	1	,000	1,014

Apéndice 9

Intervalos de confianza para las funciones de probabilidad, supervivencia e impacto

Cuando se trabaja con tablas de mortalidad (tiempo dividido en intervalos), los errores típicos (*ET*) de las funciones de *probabilidad*, *supervivencia* e *impacto* (ver Gehan 1969) vienen dados por

$$ET[\hat{f}(t_i)] = \frac{P_i q_i}{a_i} \sqrt{\sum_{j=1}^{i-1} \frac{q_j}{r_j p_j} + \frac{P_i}{r_i q_i}}, \quad \text{con } ET[\hat{f}(t_1)] = \frac{P_1 q_1}{a_1} \sqrt{\frac{P_1}{r_1 q_1}} \quad [9.25]$$

$$ET(P_i) = P_i \sqrt{\sum_{j=1}^i \frac{q_j}{r_j p_j}} \quad [9.26]$$

$$ET[\hat{h}(t_i)] = \sqrt{\frac{\hat{h}(t_i)^2}{r_i q_i} \left[1 - \left(\frac{\hat{h}(t_i) a_i}{2} \right)^2 \right]}, \quad \text{con } ET[\hat{h}(t_i)] = 0 \text{ si } q_i = 0 \quad [9.27]$$

Estos errores típicos pueden utilizarse para construir intervalos de confianza. Por ejemplo, el intervalo de confianza al 95 % de la función de supervivencia (proporción acumulada de no-eventos) en el tercer intervalo puede obtenerse de la siguiente manera:

$$SE(P_3) = 0,9659 \sqrt{\frac{0,0018}{1142,5(0,9982)} + \frac{0,0152}{984,5(0,9848)} + \frac{0,0174}{804,5(0,9826)}} = 0,00605$$

$$IC(P_3) = 0,9659 \pm 1,96(0,000605) = (0,9647, 0,9671)$$

(1,96 corresponde al cuantil 97,5 de la distribución normal tipificada). El intervalo de confianza obtenido para la función de supervivencia del tercer intervalo temporal permite concluir que la verdadera proporción acumulada de no-eventos en el tercer intervalo temporal se encuentra entre 0,9647 y 0,9671.

Cuando se utiliza el método de de Kaplan-Meier, el error típico de la función de supervivencia se estima mediante

$$SE[\hat{S}(t_i)] = \hat{S}(t_i) \sqrt{\sum_{i=1}^i \frac{q_i}{r_i - d_i}} \quad [9.28]$$

Por ejemplo, al aplicar esta ecuación a los datos de la Tabla 9.1 se obtiene, para el tercer caso (tiempo de espera 13), el siguiente error típico:

$$SE[\hat{S}(t_3)] = 0,700 \sqrt{\frac{0,100}{10-1} + \frac{0,111}{9-1} + \frac{0,125}{8-1}} = 0,145$$

Este error típico puede utilizarse para obtener intervalos de confianza para los valores estimados de la función de supervivencia:

$$IC[S(t_i)] = \hat{S}(t_i) \pm Z_{1-\alpha/2} SE[\hat{S}(t_i)] \quad [9.29]$$

donde $Z_{1-\alpha/2}$ se refiere al cuantil $1 - \alpha/2$ de la distribución normal tipificada. Por ejemplo, el intervalo de confianza de la función de supervivencia correspondiente al tercer caso de la Tabla 9.1, con una confianza del 95%, vale

$$IC[S(t_3)] = \hat{S}(t_3) \pm Z_{0,975} SE[\hat{S}(t_3)] = 0,700 \pm 1,96(0,145) = (0,984; 0,416)$$

Este intervalo indica que al cabo de 13 semanas (tiempo de espera correspondiente al tercer caso) cabe esperar, con una confianza del 95%, que la proporción de no-eventos (la proporción de pacientes en los que todavía no ha remitido el tumor) se encuentre entre 0,416 y 0,984.

Estadístico de Wilcoxon-Gehan

Para obtener este estadístico se comienza ordenando los tiempos de espera de forma ascendente (con independencia del grupo al que pertenezcan); en caso de empate se considera que los tiempos de espera de los casos no censurados son menores que los de los casos censurados). Cuando los tiempos de espera están agrupados en intervalos, a todos los casos de un mismo intervalo se les asigna el límite inferior de su intervalo. Tras ordenar los tiempos de espera se calcula:

NMI_i = número de casos no censurados cuyo tiempo de espera es menor o igual que el del i -ésimo caso.

CMI_i = número de casos censurados cuyo tiempo de espera es menor o igual que el del i -ésimo caso.

NI_i = número de casos no censurados cuyo tiempo de espera es igual que el del i -ésimo caso.

CI_i = número de casos censurados cuyo tiempo de espera es igual que el del i -ésimo caso.

A partir de aquí se obtienen las puntuaciones X_{ij} para cada sujeto (i se refiere a los sujetos y j a los grupos: $i = 1, 2, \dots, n$; $j = 1, 2, \dots, g$; g es el número de grupos). Para los casos censurados: $X_{ij} = NMI_i$. Para los no censurados: $X_{ij} = A_1 - A_2 - A_3$, donde: $A_1 = NMI_i - NI_i$, $A_2 = n_0 - CMI_i - CI_i$ y $A_3 = n_1 - NMI_i$ (donde n_0 se refiere al número de casos censurados y n_1 al número de casos no censurados).

Una vez que se tienen las puntuaciones X_{ij} de cada caso, el estadístico de Wilcoxon-Gehan ($W-G$) se calcula de la siguiente manera:

$$W-G = \frac{(n-1)}{T^2} \sum_j \frac{T_j^2}{n_j} \quad (\text{con } T_j = \sum_i^{n_j} X_{ij} \text{ y } T = \sum_j \sum_i^{n_j} X_{ij}) \quad [9.30]$$

Bajo la hipótesis de que todos los grupos poseen la misma función de supervivencia, el estadístico $W-G$ se aproxima a la distribución χ^2 con $g - 1$ grados de libertad conforme el tamaño de los grupos va aumentando.

Referencias bibliográficas

- Abad F, Olea J, Ponsoda V y García C (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis.
- Agresti A (1990). *Categorical data analysis*. New York: Wiley.
- Agresti A (2002). *Categorical data analysis* (2ª ed). New York: Wiley.
- Agresti A (2007). *An introduction to categorical data analysis* (2ª ed). New York: Wiley.
- Agresti A (2010). *Analysis of ordinal categorical data* (2ª ed). New York: Wiley.
- Agresti A y Yang M (1987). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics and Data Analysis*, 5, 9-21.
- Aitkin MA, Francis BJ y Hinde JP (2005). *Statistical modeling in GLIM 4* (2ª ed). Oxford: Oxford University Press.
- Akaike H (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, 19, 716-723.
- Amón J (1984). *Estadística para psicólogos. Probabilidad y estadística inferencial* (3ª ed). Madrid: Pirámide.
- Ato M, Losilla JM, Navarro JB, Palmer A y Rodrigo MF (2005). *Modelo lineal generalizado*. Girona: Edicions a Petició.
- Bell BA, Ferron JM y Kromrey JD (2008). Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models. *JSM Survey Research Methods*, 1122-1129.
- Bell BA, Morgan GB, Kromrey JD y Ferron JM (2010). The impact of small cluster size on multilevel models: A Monte Carlo examination of two-level models with binary and continuous predictors. *JSM Survey Research Methods*, 4057-4067.
- Bell BA, Morgan GB, Schoeneberger JA y Loudermilk BL (2010). Dancing the sample size limbo with mixed models: How low can you go? *SAS Global Forum 2010* (paper 197).
- Bickel R (2007). *Multilevel analysis for applied research. It's just regression!* New York: The Guilford Press.
- Bishop YMM (1969). Full contingency tables, logits and split contingency tables. *Biometrics*, 25, 383-399.
- Bishop YMM y Fienberg SE (1969). Incomplete two-dimensional contingency tables. *Biometrics*, 25, 119-128.
- Bishop YMM, Fienberg SE y Holland PW (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: The MIT Press.
- Blossfeld HP, Hamerle A y Mayer KU (1989). *Event history analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

- Bonett DG y Bentler PM (1983). Goodness-of-fit procedures for the evaluation and selection of log-linear models. *Psychological Bulletin*, 93, 149-166.
- Bozdogan H (1987). Model selection and Akaike's selection criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345-370.
- Breslow NE (1970). A generalized Kruskal-Wallis test for comparing K samples subject to unequal pattern of censorship. *Biometrika*, 57, 579-594.
- Brown H y Prescott R (1999). *Applied mixed models in medicine*. New York: Wiley.
- Cameron AC y Trivedi PK (1998). *Regression analysis of count data*. Cambridge: Cambridge University Press.
- Cnaan A, Laird NM y Slator P (1997). Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, 16, 2349-2380.
- Clogg CC y Shihadeh ES (1994). *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage.
- Collett D (1994). *Modelling survival data in medical research*. London: Chapman and Hall.
- Corbeil RR y Searle SR (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed models. *Technometrics*, 18, 31-38.
- Cox DR (1959). The analysis of exponentially distributed life-times with two types of failures. *Journal of the Royal Statistical Society, B*, 21, 411-421.
- Cox DR (1970). *The analysis of binary data*. London: Chapman and Hall.
- Cox DR (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, B*, 34, 187-220.
- Cox DR y Oakes D (1984). *Analysis of survival data*. London: Chapman and Hall.
- Cox DR y Snell EJ (1968). A general definition of residuals. *Journal of the Royal Statistical Society, B*, 30, 248-275.
- Deming WE y Stephan FF (1940). On the least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.
- Dunteman GH y Ho MHR (2006). *Introduction to generalized linear models*. Thousand Oaks, CA: Sage.
- Fienberg SE (1970). The analysis of multidimensional contingency tables. *Ecology*, 51, 419-433.
- Fienberg SE (1972). The analysis of incomplete multiway contingency tables. *Biometrics*, 28, 177-202.
- Fienberg SE (1980). *The analysis of cross-classified categorical data* (2ª ed). Cambridge, MA: The MIT Press.
- Fisher RA (1922). On the interpretation of chi-square from contingency tables, and the calculation of P . *Journal of the Royal Statistical Society*, 85, 87-94.
- Fisher RA (1924). The conditions under which X^2 measures the discrepancy between observation and hypothesis. *Journal of the Royal Statistical Society*, 87, 442-450.
- Fisher RA (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700-725.
- Fisher RA (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society, A*, 144, 285-307.
- Fox J (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.
- Gardner W, Mulvey EP y Shaw EC (1995). Regression analysis of counts and rates: Poisson, over-dispersed Poisson, and negative binomial models. *Psychological Bulletin*, 118, 392-404.
- Gehan EA (1965a). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52, 203-223.
- Gehan EA (1965b). A generalized two-sample Wilcoxon test for doubly-censored data. *Biometrika*, 52, 650-653.
- Gehan EA (1969). Estimating survival function from the life table. *Journal of Chronic Diseases*, 21, 629-644.
- Gill J (2001). *Generalized linear models. A unified approach*. Thousand Oaks, CA: Sage.

- Goldstein H (2003). *Multilevel statistical models* (3ª ed). New York: Halstead Press.
- Goodman LA (1968). The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing data. *Journal of the American Statistical Association*, 63, 1091-1131.
- Goodman LA (1970). The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association*, 65, 226-256.
- Goodman LA (1971). The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classification. *Technometrics*, 13, 33-61.
- Grambsch PM y Therneau TM (1994). Proportional hazards tests in diagnostics based on weighted residuals. *Biometrika*, 81, 515-526.
- Green PJ (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). *Journal of the Royal Statistical Society, B*, 46, 149-192.
- Gross, AJ y Clark, VA (1975). *Survival distributions: Reliability applications in the biomedical sciences*. New York: Wiley.
- Haberman SJ (1973). The analysis of residuals in cross-classification tables. *Biometrics*, 29, 205-220.
- Haberman SJ (1974). *The analysis of frequency data*. Chicago: University of Chicago Press.
- Haberman SJ (1978). *Analysis of qualitative data: Introductory topics*. New York: Academic Press.
- Haberman SJ (1979). *Analysis of qualitative data: New developments*. New York: Academic Press.
- Haberman SJ (1982). The analysis of dispersion of multinomial responses. *Journal of the American Statistical Association*, 77, 568-580.
- Hanley JA y McNeil BJ (1982). The meaning and use of the area under receiver operating characteristic (ROC). *Radiology*, 143, 29-36.
- Harrell FE (2001). *Regression modeling strategies with applications to linear models, logistic regression and survival analysis*. New York: Springer.
- Hauck WW y Donner A (1977). Wlad's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72, 851-853.
- Heck RH y Thomas SL (2000). *An introduction to multilevel modeling techniques*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Henderson DA y Denison DN (1984) Stepwise regression in social and psychological research. *Psychological Reports*, 64, 261-267.
- Hess KR (1995). Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Statistics in Medicine*, 14, 1707-1723.
- Hosmer DW, Hosmer T, Le Cessie S y Lemeshow S (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16, 965-980.
- Hosmer DW y Lemeshow S (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10, 1043-1069.
- Hosmer DW y Lemeshow S (1999). *Applied survival analysis: Regression modeling of time to event data*. New York: Wiley.
- Hosmer DW y Lemeshow S (2000). *Applied logistic regression* (2ª ed). New York: Wiley.
- Hox J (2010). *Multilevel analysis. Techniques and applications* (2ª ed). New York: Routledge.
- Huberty CJ (1989). Problems with stepwise methods: Better alternatives. En B. Thompson (Ed), *Advances in social science methodology* (vol 1, pp. 43-70). Greenwich, CT: JAI Press.
- Hutcheson G y Sofroniou N (1999). *The multivariate social scientist. Introductory statistics using generalized linear models*. London: Sage.
- Hurvich CM y Tsai CL (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297-307.
- Irwin JO (1949). The standard error of an estimate of expectational life. *Journal of Hygiene*, 47, 188-189.
- Jaccard J (2001). *Interaction effects in logistic regression*. Thousand Oaks, CA: Sage.

- Jaccard J y Turrissi R (2003). *Interaction effects in multiple regression* (2ª ed). Thousand Oaks, CA: Sage.
- Jennings DE (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, 81, 471-476.
- Jennrich RI y Sampson PF (1976). Newton-Raphson and related algorithms for maximum likelihood variance components estimation. *Technometrics*, 18, 11-17.
- Judd ChM, McClelland GH y Ryan CS (2009). *Data analysis. A model comparison approach* (2ª ed). New York: Routledge.
- Kalbfleisch JD y Prentice RL (1980). *The statistical analysis of failure time data*. New York: Wiley.
- Kaplan EL y Meier P (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457-481.
- Keppel G y Wickens ThD (2004). *Design and analysis. A researcher's handbook* (4ª ed). Englewood Cliffs, NJ: Prentice-Hall.
- Koehler KJ (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. *Journal of the American Statistical Association*, 81, 483-493.
- Koehler KJ y Larntz K (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, 75, 336-344.
- Kleinbaum DG y Klein M (2002). *Logistic regression: A self-learning text*. New York: Springer.
- Kutner MH, Nachtsheim CJ, Neter J y Li W (2005). *Applied linear statistical models* (5ª ed). McGraw-Hill/Irwin.
- Lawless JF (1982). *Statistical models and methods for lifetime data*. New York: Wiley.
- Lawless JF y Singhal K (1978). Efficient screening of nonnormal regression models. *Biometrics*, 34, 318-327.
- Lee ET (1992). *Statistical methods for survival data analysis* (2ª ed). New York: Wiley.
- Long JS (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Longford NT (1993). *Random coefficient models*. New York: Oxford University Press.
- Luke DA (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.
- Maas C y Hox J (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, 58, 127-137.
- Maas C y Hox J (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1, 86-92.
- Magidson J (1981). Qualitative variance, entropy, and correlation ratios for nominal dependent variables. *Social Science Research*, 10, 177-194.
- Mantel N (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50, 163-170.
- Mantel N (1970). Incomplete contingency tables. *Biometrics*, 26, 291-304.
- Maxwell SE y Delaney HD (2004). *Designing experiments and analyzing data. A model comparison perspective* (2ª ed). Mahwah, NJ: Lawrence Erlbaum Associates.
- McCullagh P (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, B*, 42, 109-142.
- McCullagh P y Nelder JA (1989). *Generalized linear models* (2ª ed). New York: Chapman and Hall.
- McCulloch CE y Searle SR (2001). *Generalized, linear, and mixed models*. New York: Wiley.
- McFadden D (1974). Conditional logit analysis of qualitative choice behavior. En P Zarembka (Ed), *Frontiers in econometrics* (pp 105-142). New York: Academic Press.
- Menard S (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54, 17-24.
- Menard S (2001). *Applied logistic regression analysis* (2ª ed). Thousand Oaks, CA: Sage.
- Montgomery DC, Peck EA y Vining GG (2001). *Introduction to linear regression analysis* (3ª ed). New York: Wiley.
- Nagelkerke NJD (1991). A note on the general definition of the coefficient of determination. *Biometrika*, 78, 691-692.

- Nelder JA y Wedderburn, RWM (1972). Generalized linear models. *Journal of the Royal Statistical Society, A*, 135, 370-384.
- Neyman J y Pearson ES (1928). On the use and interpretation of certain test criteria for purposes of statistical inference (2ª parte). *Biometrika*, 20, 263-294.
- Pampel FC (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage.
- Pardo A (2002). *Análisis de datos categóricos*. Madrid: Ediciones de la Universidad Nacional de Educación a Distancia.
- Pardo A, Ruiz MA y San Martín R (2009). *Análisis de datos en ciencias sociales y de la salud* (vol 1). Madrid: Síntesis.
- Pardo A y San Martín R (1994). *Análisis de datos en psicología II*. Madrid: Pirámide.
- Pardo A y San Martín R (1998). *Análisis de datos en psicología II* (2ª ed). Madrid: Pirámide.
- Pardo A y San Martín R (2010). *Análisis de datos en ciencias sociales y de la salud* (vol 2). Madrid: Síntesis.
- Parmar MK y Machin D (1995). *Survival analysis: A practical approach*. New York: Wiley.
- Pearson K (1911). On the probability that two independent distributions of frequency are really samples from the same population. *Biometrika*, 8, 250-254.
- Peto R y Peto J (1972). Asymptotically efficient rank invariant procedures. *Journal of the Royal Statistical Society, A*, 135, 185-207.
- Pierce DA y Schafer DW (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*, 81, 977-983.
- Powers DA y Xie Y (1999). *Statistical methods for categorical data analysis*. San Diego, CA: Academic Press.
- Pregibon D (1981). Logistic regression diagnostics. *Annals of Statistics*, 9, 705-724.
- Prentice RL y Marek P (1979). A quantitative discrepancy between censored data rank tests. *Biometrics*, 35, 861-867.
- Raftery AE (1995). Bayesian model selection in social research. In PV Marsden (Ed) *Sociological Methodology* (pp 111-163). London: Tavistock.
- Rao CR (1973). *Linear statistical inference and its application* (2ª ed). New York: Wiley.
- Rao CR y Kleffe J (1988). *Estimation of variance components and applications*. Amsterdam: North-Holland.
- Raudenbush SW (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, 52, 501-525.
- Raudenbush SW (2008). Many small groups. En J de Leeu y E Meijer (2008): *Handbook of multilevel analysis* (pp 207-236). New York: Springer
- Raudenbush SW y Bryk AS (2002). *Hierarchical linear models: Applications and data analysis methods* (2ª ed). Thousand Oaks, CA: Sage.
- Raudenbush SW, Spybrook J, Congdon R, Liu X y Martínez A (2011). Optimal Design software for multilevel and longitudinal research (versión 3.01). Disponible en "<http://www.wtgrantfdn.org>" dentro del apartado "resources" en la opción "consultation-service-and-optimal-design".
- Ríos S (1977) *Métodos estadísticos* (2ª ed). Madrid: Ediciones del Castillo.
- Scherbaum CA y Ferreter JM (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, 12, 347-367.
- Schoenfeld D (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69, 239-241.
- Schwarz, G (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Searle SR, Casella G y McCulloch CE (1992). *Variance components*, New York: Wiley.
- Singer J y Willett J (2003). *Applied longitudinal data analysis*. New York: Oxford University Press.
- Snijders TAB y Bosker RJ. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, 18, 237-259.
- Snijders, TAB y Bosker, RJ (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

- Stevens JP (1992). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Tabachnick BG y Fidell LS (2001). *Using multivariate statistics* (4ª ed). Boston: Allyn and Bacon.
- Tarone RE y Ware J (1977). On distribution free tests of the equality of survival distributions. *Biometrika*, 64, 156-160.
- Theil H (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology*, 76, 103-154.
- Therneau TM, Grambsch PM y Fleming TR (1990). Martingale-based residuals for survival models. *Biometrika*, 77, 147-160.
- Twisk JWR (2006). *Applied multilevel analysis. A practical guide*. Cambridge: Cambridge University Press.
- Verbeke G y Molenberghs G (2000). *Linear mixed models for longitudinal data*. New York: Springer.
- Wickens ThD (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Wilks, SS (1935). The likelihood test of independence in contingency tables. *Annals of Mathematical Statistics*, 6, 190-196.

Índice de materias

A

- Ajuste (de un modelo lineal), 29-32
 - global, 29-32, 119-120
 - contribución de cada variable independiente al, 32, 63, 170, 181, 183, 211, 225-226, 252, 279
 - criterio de máximo ajuste, 28, 118, 197, 278
 - criterio de parsimonia, 28, 32, 118, 197, 278
 - por pasos, 28, 64, 170, 197-202, 228, 278, 282-291, 375
 - porcentaje de casos clasificados correctamente, 32, 169, 173-176, 179, 182, 186-187, 227
 - significación estadística, 30-31, 47-49, 54, 61, 63, 67-68, 84-85, 90-91, 170-172, 180-181, 199-200, 218, 222-223, 231-232, 244-245, 251-252, 260-262, 277-278, 293-294, 301, 304, 306, 308-310, 325-327, 371
 - significación sustantiva, 31-32, 47-49, 63, 67, 172-173, 182, 219, 222, 232, 242, 246, 200, 327-328
 - valoración del cambio en el ajuste asociado a un único caso, 211
- Ajuste proporcional iterativo, 277
- Akaike, criterios de información *AIC*, *AICC*, *CAIC*, 84, 119-120, 261-262
- Aleatorio, componente (ver *componentes de un modelo lineal*)
- Aleatorios, efectos, 77-78
- Análisis de correlación canónica, 19, 26
- Análisis de covarianza:
 - ajuste, 54, 58, 61
 - estimaciones de los parámetros, 53, 59-60
 - lógica del, 51
 - modelos:
 - dos factores con medidas repetidas en uno, 113-116
 - un factor, efectos aleatorios, 134-136
 - un factor, efectos fijos, 52-62
 - pendientes heterogéneas, 62
 - pronósticos, 53
 - supuestos, 54-57
- Análisis de regresión de Cox (ver *Cox, regresión de*)
- Análisis de regresión de Poisson (ver *Poisson, regresión de*)
- Análisis de regresión lineal (ver *regresión lineal*)
- Análisis de regresión logística (ver *regresión logística*)
- Análisis de regresión multinivel (ver *multinivel, modelos lineales*)
- Análisis de regresión nominal (ver *regresión nominal*)
- Análisis de regresión ordinal (ver *regresión ordinal*)
- Análisis de supervivencia (ver *supervivencia, análisis de*)
- Análisis de varianza, 23, 43-49
 - ajuste, 47-49
 - estimaciones, 46
 - modelos:
 - dos factores con medidas repetidas en ambos, 102-104
 - dos factores con medidas repetidas en uno, 104-113
 - dos factores, efectos mixtos, 88-94
 - un factor, efectos fijos, 45-46
 - un factor, efectos aleatorios, 80-87, 129-131
 - un factor, medidas repetidas, 97-102, 147-150
 - pronósticos, 46-47
 - supuestos, 49
- Análisis de varianza con medidas repetidas (enfoque mixto):
 - comparaciones múltiples, 101-102, 108
 - efecto de la interacción, 112-113
 - efectos simples, 109-112
 - estructura de los datos, 95-97
 - modelo de dos factores con medidas repetidas en ambos, 102-104
 - modelo de dos factores con medidas repetidas en uno, 104-113
 - modelo de un factor, 97-102
- Análisis de varianza multivariado, 19

Análisis loglineal (ver *loglineales, modelos*)

Asociación:

completa, 268

en tablas de contingencias, 267-269

parcial, 268

homogénea, 268

Atípicos, casos, 33-34, 208-211, 376-377

B

Bayesiano, criterio de información *BIC*, 84, 119-120, 261-262

Binomial, distribución, 24, 36-37, 40-41, 165, 206, 332-333

Binomial negativa, distribución, 35, 262-263

Bloques aleatorios, 50

Bondad de ajuste (ver *ajuste*):

Bonferroni, corrección para comparaciones múltiples, 101, 105, 108, 110, 362

C

Canónica, correlación, 19, 26

Canónico, parámetro, 25, 35-36

Casos atípicos, 33-34, 208-211, 376-377

Casos influyentes, 34, 211-212, 377-378

Ceros estructurales *o a priori*, 300

Ceros muestrales, 299

Chi-cuadrado (ver *ji-cuadrado*)

Clasificación, tabla de, 175, 179, 185-186, 227 (ver también *matriz de confusión*)

Coefficiente de determinación, 31, 67, 172-173, 378

Coefficiente de incertidumbre, 328

Coefficiente de variación, 83

Coefficientes de regresión, 21, 24, 31, 51, 69-72, 124-129, 132-133, 135-138, 142-144, 147, 150-155, 161, 163, 171, 176-177, 182-185, 192-197, 216, 220-221, 224-225, 230, 232-235, 241, 246-259, 317-320, 367-368, 374, 378, 382

Colinealidad, 204-205

Comparaciones múltiples, 48, 59, 62, 92-94, 98, 101-102, 104, 108-113, 314-316, 352-353, 361-366

Componentes de un modelo lineal, 24-26

componente aleatorio, 24, 33, 35-36, 44, 52, 129, 243-244, 260, 270, 293

componente sistemático, 24-25, 27, 35, 51, 271

función de enlace, 25-26, 29, 43-44, 165, 213, 216, 229, 236, 244, 261, 270, 368

Concentración, índice de, 326-328

Contingencias, tabla de (ver *tabla de contingencias*)

Contrastes (polinómicos, especiales), 188, 203-204

Cook, distancia de, 211-212

Correlación intraclase, 86-87, 100-101, 118-119, 130-131, 134, 157

Covarianza, 118, 149, 154

Covarianza, análisis de (ver *análisis de covarianza*)

Cox, regresión de, 366-383

ajuste global, 371

casos atípicos e influyentes, 376-378

covariables categóricas, 373-375

covariables dependientes del tiempo, 378-383

impacto proporcional, 368-369

impacto relativo (razón de impactos), 367

modelo, 367-368

por pasos, 375

residuos de Cox y Snell, 376

residuos de martingala, 376

residuos de Schoenfeld, 376-377

residuos parciales, 376-377

supuestos, 375-376

Cox y Snell:

R^2 de, 173, 182

residuos de, 376

Criterios de información, 84, 119-120, 261-262

Cuasi-independencia, modelo de, 281, 301-304

Curva COR (curva característica de operación del receptor), 186

Curvas de crecimiento, 146-155, 158

D

Desajuste, 31-32, 47-48, 61,

máximo, 31, 47, 66, 171,

aumento en el, 48-49, 54, 67

reducción en el, 67, 172, 200, 218-219, 222-223, 226, 231-232, 245-246, 251-252,

Desvianza, 30-31, 47-49, 54, 61, 66-67, 84, 171-172, 207, 218, 222-223, 231, 245, 261-262, 370

Desvianza, residuos de, 209-210, 280

Dispersión, parámetro de (ver *parámetro de escala*)

Distancia de Cook, 211-212

Dunn-Bonferroni, prueba para comparaciones múltiples, 101, 108, 110, 362

E

Efecto del diseño, 157-158

Efectos:

fijos, aleatorios, mixtos, 77-78

interacción (ver *interacción entre variables independientes*)

simples, 105, 109-112

Enlace, función de (ver *componentes de un modelo lineal*)

Entropía, índice de, 326-328
 Error (en un modelo lineal), 21, 22, 24, 31, 33, 45, 47, 49, 50-51, 53, 64, 66, 74-75, 79, 81, 89, 97, 102-103, 114, 116, 120-121, 124, 126, 128-129, 137, 141-142, 148, 151, 162, 172, 206, 209, 242-243
 Error de especificación, 203
 Escala, parámetro de (ver *parámetro de escala*)
 Esquemas de muestreo (multinomial, Poisson, multinomial condicional), 331-334
 Estadísticos mínimo-suficientes, 334-335
 Exponencial, familia, 35-38

F

F , estadístico de Fisher, 31, 48-49, 54, 61, 67-68
 Factores de inflación de la varianza, 205
 Fijos, efectos, 77-78
 Fisher:
 familia exponencial, 35-38
 máxima verosimilitud, 38-41
 método *scoring*, 28, 41, 122
 Fuentes de variabilidad (ver *variabilidad*)
 Función de enlace (ver *componentes de un modelo lineal*)

G

G^2 (ver *razón de verosimilitudes*)
 General, modelo (ver *modelo lineal general*)
 Generalizado, modelo (ver *modelo lineal generalizado*)

H

Homocedasticidad o igualdad de varianzas (ver *suposuestos de un modelo lineal*)

I

Impacto proporcional, 368-369
 Impacto relativo (razón de impactos), 367
 Incertidumbre, coeficiente de, 328
 Independencia
 completa, 268, 272, 276
 condicional, 268, 276, 288-289, 298, 322
 entre observaciones, 123
 hipótesis de, 168, 217, 261, 332-332
 modelo de, 269-270, 280, 312, 330, 335
 supuesto de, 33, 49, 55, 56, 68, 79, 161, 205-206
 Inflación de la varianza, factores de, 205
 Influencia, valor de (*leverage*), 34, 210-211

Influyentes, casos, 34, 211-212, 377-378
 Información, criterios de, 261-262
 Intergrupos, variabilidad (ver *variabilidad*)
 Interacción entre variables independientes, 25, 57, 62, 68-73, 88, 93, 102, 107-108, 112-113, 116, 141, 143-144, 153-154, 190-197, 227, 235, 254-258, 382
 Intersujetos, variabilidad (ver *variabilidad*)
 Intersujetos, factor, 104, 114-116
 Intraclass, coeficiente de correlación, 86-87, 100-101, 130-131, 134
 Intrasujetos, factor, 98, 114-115, 117, 153
 Intrasujetos, variabilidad (ver *variabilidad*)
 Intragrupos o error, variabilidad (ver *variabilidad*)

J

Ji-cuadrado, estadístico de Pearson, 223, 261, 277
 Jerarquía, principio de, 274-275
 Jerárquica, estructura, 123
 Jerárquica, regresión (ver *regresión por pasos*)
 Jerárquicos, modelos, 31, 198, 269-291

K

Kaplan-Meier, método de, 354-366

L

Lineal, modelo (ver *modelos lineales*)
 Lineal, relación, 22, 51, 55, 161-162, 203-204, 240-242, 247, 294-295
 Linealidad (supuesto del análisis de regresión lineal), 33, 68, 161, 203-204
 Logística (ver *regresión logística*)
 Logit, modelos, 316-331
 ajuste global, 325-328
 correspondencia entre los modelos logit y los loglineales, 320-324
 medidas de entropía y concentración, 327-328
 modelos, 317-319
 parámetros, 318-320, 328-331
 significación estadística, 325-327
 significación sustantiva, 327-328
 Loglineales, modelos, 265-665
 ajuste en un único paso, 292-296
 ajuste por pasos, 282-291
 ajuste proporcional iterativo, 277
 asociación en tablas de contingencias, 267-269
 casillas vacías, 298-300
 ceros estructurales o *a priori*, 300

ceros muestrales, 299
 comparaciones entre niveles, 314-316
 descripción general, 265-266
 esquemas de muestreo (multinomial, multinomial condicional, Poisson), 331-334
 estadísticos de ajuste, 277-278
 estadísticos mínimo-suficientes, 334-335
 estimaciones de las frecuencias esperadas, 276-277
 estimaciones de los parámetros, 296-298
 grados de libertad, 335
 gráficos de los residuos, 295
 modelo de cuasi-independencia, 301-304
 modelo de dependencia, 270-271
 modelo de independencia, 269-270
 modelo de simetría completa, 304-307
 modelo de simetría relativa, 307-310
 modelo saturado, 272, 273-274
 modelos generales, 292-316
 modelos jerárquicos, 269-291
 modelos no comprensivos, 272
 modelos para tasas de respuesta, 310-314
 notación en tablas de contingencias, 266-267
odds ratio generalizada, 316
 parámetros independientes, 271-273
 principio de jerarquía, 274-275
 procedimientos SPSS, 281
 residuos, 279-281
 selección del mejor modelo, 278-279
 símbolos y configuraciones, 275
 tablas cuadradas, 300-310
 tablas incompletas, 298-300

M

Martingala, residuos de, 376

Matriz:

autorregresiva, 118-120
 de confusión, 169, 174, 182, 227
 de varianzas-covarianzas, 82, 89, 94, 97, 99-100, 104, 114, 116-121
 del diseño, 120-121, 296
 diagonal, 81-82, 89, 117, 120-121
G, 82, 84, 89, 121, 138-140
R, 81-82, 89, 97, 99-100, 102, 116-119, 121
 simetría compuesta, 100, 102, 116-120, 140
 sin estructura, 117-118, 120, 145
 Toeplitz, 119-120

Máxima verosimilitud:

estimación por, 28, 30, 39-41, 122, 171, 276, 334-335, 356, 369
 función de, 38-39

Medias como resultados, 131-134

Medias cuadráticas, 86, 103

Medias estimadas, 93-94, 101-102, 108-109

Medias y pendientes como resultados, 140-146

Mínimo-suficientes, estadísticos, 276-277, 334-335

Mínimos cuadrados, 28, 30, 46, 53, 64, 121

Mínimos cuadrados ponderados iterativamente, 28, 41

Mixtos, efectos, 77-78

Mixtos, modelos lineales (ver *modelos lineales mixtos*)

Modelo, 20

Modelos lineales, 20-41

aditivos, 68

cómo ajustarlos, 27-34

componentes de los, 24-26

clasificación de los, 26-27

no aditivos, 69

qué son los, 20-23

Modelos lineales generales o clásicos, 23, 31, 33, 26, 44, 73-74, 78, 123, 43-75, 78, 82, 126, 161, 207, 241, 367

Modelos lineales generalizados, 26, 28, 33, 165, 216, 229, 368, 213, 244, 270

Modelos lineales mixtos, 26, 28, 33, 38, 77-122, 123, 128-129, 137, 150

Modelos lineales multinivel, 26, 79, 123-158, 198

Multicolinealidad (ver *colinealidad*)

Multinivel, modelos lineales, 123-158

curvas de crecimiento, 146-155, 158

estructuras jerárquicas o multinivel, 123-124, 146

modelo de coeficientes aleatorios, 136-140

modelo de medias como resultados, 131-134

modelo de medias y pendientes como resultados, 140-146

modelo del nivel 1, modelo del nivel 2 y modelo mixto o combinado, 128, 129, 132, 135, 137, 147, 150

modelo incondicional o nulo, 129-131

modelos de medidas repetidas (curvas de crecimiento), 146-155

coeficientes aleatorios, 147-150

medias-pendientes como resultados, 150-155

qué es un modelo multinivel, 124-129

tamaño muestral, 155-158

un factor, efectos aleatorios, 129-131

N

Nagelkerke, R^2 de, 173, 182, 200, 219, 222, 232

Nominal, regresión (ver *regresión nominal*)

Normalidad:

gráficos de, 295

supuesto de, 33, 49, 55-56, 68, 161, 243, 247, 295

O

Odds, 25, 164-166, 170, 177-178, 184-185, 192-197, 217, 220, 225, 230, 233, 235, 318-319, 329
Odds proporcionales, 235-236
Odds ratio, 168, 178, 184-185, 190, 193-197, 213, 217, 220-221, 231, 314
Odds ratio generalizada, 316
 Ordinal, regresión (*regresión ordinal*)

P

Parámetro de escala, 35, 37, 207-208, 228-229, 260-262
 Parciales, residuos, 376-377
 Parsimonia, criterio de, 28, 32, 118, 197, 278
 Patrones de variabilidad, 25, 37, 165, 206-208, 216, 218, 223, 239, 265, 311, 325, 335
 Pearson:
 coeficiente de correlación, 31, 240
 prueba X^2 , 223, 261, 277, 280, 299
 residuos tipificados, 209, 280, 290, 294
 Poisson, distribución, 24, 26, 37-38
 Poisson, regresión de, 239-263
 ajuste global, significación estadística, 244-245, 251-252
 ajuste global: significación sustantiva, 246
 componente aleatorio, 244
 función de enlace, 243-244
 Interacción entre variables independientes, 254-258
 interpretación de los coeficientes, 247, 253
 modelo, 243
 significación de los coeficientes, 246, 252
 sobredispersión, 260-261, 262-263
 tasas de respuesta, 258-260
Post hoc, comparaciones (ver *comparaciones múltiples*)
 Predictor lineal, 24-25, 43, 165, 243
 Principal, efecto, 107, 151, 286
 Probit, función, 162, 212
 Probit, regresión, 212-214
 Pronósticos, 23-30, 46-47, 53, 60, 64-66, 161-164, 173-174, 185-187, 212, 226-227, 248-249, 273, 318-319, 330-331

R

Razón de verosimilitudes, 30-32, 84, 90, 171-172, 176, 180, 198-199, 201, 218, 222-226, 231, 245, 251-253, 277-280, 285, 288, 290, 293, 301, 304, 306, 308, 310, 312, 325-327, 335, 371

Recuento, 239

Reducción proporcional del error, medidas de, 174

Regresión curvilínea, 162, 212

Regresión jerárquica (ver *regresión por bloques y regresión por pasos*)

Regresión lineal, 22, 63-73

 ajuste, 66-68

 estimaciones, 64

 interacción entre variables independientes, 68-73

 modelo, 63-64

 pronósticos, 65-66

 supuestos, 68

Regresión logística dicotómica o binaria, 159-214

 ajuste global, 170-173

 casos atípicos, 208-211

 casos influyentes, 208, 211-212

 clasificación, 174-176, 186-187

 coeficientes de regresión, 161, 176-178, 182-185, 192-197, 203-204

 covariables categóricas, 187-190

 factores de inflación de la varianza, 205

 función logística, 162-164

 interacción entre covariables, 190-197

 modelo, 165, 191

 por pasos, 197-202

 pronósticos, 173, 185-186

 residuos, 209-212

 significación estadística, 170-172, 180-181

 significación sustantiva, 172-173, 182,

 supuestos, 203-208

 dispersión proporcional a la media, 206-208

 independencia, 205-206

 linealidad, 203-204

 no colinealidad, 204-205

 transformación logit, 164-165

Regresión logística nominal, 215-229

 ajuste global, 218-219, 222-223

 clasificación, 227

 coeficientes de regresión, 219-221, 224-225

 interacción entre variables independientes, 227

 modelo, 216-217, 221

 pronósticos, 226

 regresión por pasos, 228

 sobredispersión, 228-229

Regresión logística ordinal, 229-237

 ajuste global, 231-232, 234

 coeficientes de regresión, 232-235

 funciones de enlace, 236-237

 interacción entre variables independientes, 235

 modelo, 230

odds proporcionales, 235-236

Regresión multinivel (ver *multinivel, modelos lineales*)

Regresión por bloques, 198-199
 Regresión por pasos, 197-202, 229 (ver *ajuste por pasos*)
 Relación lineal, 22, 51, 55, 161-162, 203-204, 240-242, 247, 294-295
 Relación monótona, 167, 212, 345
 Residuos, 33-34, 47, 54, 61, 66, 209, 279
 corregidos, 280-281, 291
 de Cox y Snell, 376
 de martingala, 376
 de desviación, 209-210, 280-281
 de Schoenfeld, 376-377
 gráficos de los, 295, 377
 parciales, 376-377
 studentizados, 211
 tipificados o de Pearson, 209-210, 280
 varianza de los, 86, 92, 99-100, 104, 108, 130, 133, 136, 139, 144, 149, 154,

S

Saturado, modelo lineal, 27, 30, 171, 223, 272-276, 279-288, 310-311, 317-318,
 Schoenfeld, residuos de, 376-377
 Significación estadística, 29-32
 Significación sustantiva, 29-32
 Simetría completa, hipótesis de, 304-307
 Simetría relativa, hipótesis de, 307-310
 Simetría compuesta, 100, 102, 116-120, 140
 Simples, efectos, 109-112
 Simpson, paradoja de, 268
 Sistemático, componente (ver *componentes de un modelo lineal*)
 Sobredispersión, 33, 207-208, 228-229, 260-261, 262-263
 Studentizados, residuos, 211
 Sumas de cuadrados, 30, 31, 49, 61, 66, 170-171
 Supervivencia, análisis de, 337-385
 Breslow, estadístico de, 361-366
 caso censurado, 338, 340-342, 350, 355
 cómo comparar tiempos de espera, 352-354, 361-366
 errores típicos de las funciones de supervivencia e impacto, 383-384
 estadístico producto-límite, 355-357
 evento terminal, 338
 gráficos de los tiempos de espera, 359-361
 impacto, función de, 346-347, 383-384, 360
 impacto, tasa de, 367
 Kaplan-Meier, método de, 354-366
 log-rango, estadístico, 361-366
 media de los tiempos de espera, 356

mediana de los tiempos de espera, 343, 351, 356
 regresión de Cox (ver *regresión de Cox*)
 supervivencia, función de, 345-346, 383-384, 359-360
 supervivencia, tiempo de, 338
 tablas de mortalidad, 340-354
 Tarone-Ware, estadístico de, 361-366
 Supuestos de un modelo lineal, 32-33
 dispersión igual a la media, 33, 207-208, 228-229, 260-261, 262-263
 homocedasticidad o igualdad de varianzas, 26, 33, 49, 55-56, 68, 79, 161
 independencia, 33, 49, 55-56, 68, 79, 101, 123, 161, 205-206
 linealidad, 33, 68, 161, 203-204
 no colinealidad, 33, 68, 204-205
 normalidad, 33, 49, 55-56, 68, 161, 243, 247, 295
 simetría compuesta, 100, 102, 116-120, 140

T

Tabla de contingencias, 266-269
 Tablas cuadradas, 300-310
 Tablas incompletas, 298-300
 Tamaño muestral efectivo, 157
 Tasa de error, 101, 108, 156, 352, 362
 Tasas de respuesta (cómo analizarlas), 258-260, 310-314
 Tendencia, comparaciones de, 204
 Test (sentencia SPSS para comparaciones múltiples), 110-113
 Tipificados, residuos, 209-210, 280
 Tolerancia, nivel de, 205
 Transformación logit, 164-166, 186

U

Unidades del primer nivel, 131, 146-147
 Unidades del segundo nivel, 131, 146-147

V

Variabilidad:
 entre medias, 86, 93, 102, 104, 108, 126-128, 130, 132, 137, 139, 141, 145, 148-149, 151, 154
 entre pendientes, 126-128, 137, 139, 141, 145-148, 150-151, 155
 estimación ponderada por la, 82
 explicada, 86-88, 104, 108, 131, 134, 145-146, 173

- intergrupos, 74, 130
- intermedidas, 100
- intersujetos, 97, 100, 102, 104, 108, 114, 149, 151
- intragrupos o error, 50-51, 74, 86, 92-93, 97, 128, 130, 139, 141, 144
- intrasujetos, 97, 100, 104, 108, 148-149, 154
- nivel 1, 131, 133-137, 139, 148-149, 151, 154
- nivel 2, 126, 131-134, 137, 145, 149
- no explicada, 21, 100, 104
- patrones de, 25, 37, 165, 206-208, 216, 218, 223, 239, 265, 311, 325, 335
- total, 86, 97, 104, 108, 131
- Variable:
 - centrada, 69, 124, 132, 194, 196, 247, 251, 256
 - covariable, 51, 159
 - dependiente o respuesta, 20
 - dummy* (ficticia, indicador), 188
 - extraña, 50
 - independiente o predictora, 20
- Variación, coeficiente de, 83
- Varianza, análisis de (ver *análisis de varianza*)
- Varianza común o explicada, 30, 86-87, 100, 131, 134, 327-328 (ver *variabilidad explicada*)
- Varianza no explicada, 100, 104, 120, 327
- Varianzas-covarianzas, matriz de, 82, 89, 94, 97, 99-100, 104, 114, 116-121
- W**
 - Wald, estadístico de, 85, 87, 133-134, 176, 183, 225, 246, 372
 - Wilcoxon-Gehan, estadístico de, 352-353, 384-385